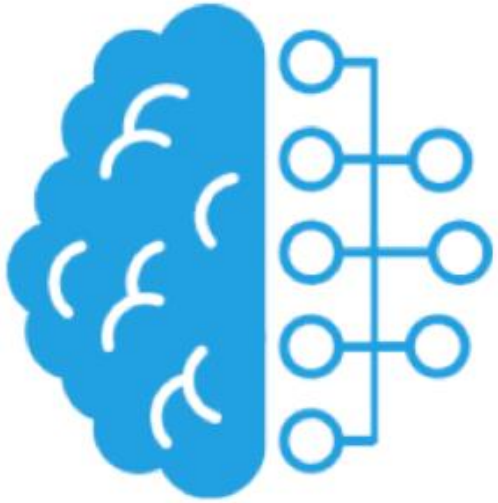
# Aprendizaje de Máquina
## Clase 1

Claudio Delrieux – DIEC - UNS
cad@uns.edu.ar



# Machine Learning: Introducción

ML es probablemente la primera en surgir de las múltiples ramas de la Inteligencia Artificial.

Su propósito es obtener sistemas que tengan la capacidad de aprender a realizar alguna tarea o cumplir algún propósito sin requerir una programación explícita.

En general, el propósito primario del uso de ML es aprender a realizar predicciones basadas en los datos.

# Machine Learning: Introducción

Las aplicaciones de ML cubren todo el rango de tareas para las cuales una programación estática sería poco práctica:

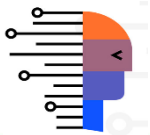
- No hay expertos humanos que sepan resolver la tarea.
- Hay expertos humanos, pero éstos no sabrían cómo describir su accionar.
- El contexto es abierto o se modifica.
- Hay muchas instancias (usuarios) de una misma tarea, cada una con restricciones locales diferentes.
- La información necesaria/disponible es copiosa/cambiante.

# Machine Learning: Introducción

En general, el ML se basa en plantear un «modelo» del dataset (el cual es explícito o implícito, y puede o no ser determinado bajo supervisión).

ML está estrechamente relacionada con los analíticos, en particular con los predictivos (aunque los diagnósticos y los prescriptivos también requieren ML).

También la relación es muy grande con el reconocimiento de patrones y con la minería de datos.



# Machine Learning: Introducción

Otros campos de estudio se enfocan en esta misma problemática:

**Estadística**: Se enfoca en el entendimiento de propiedades (paramétricas) de los fenómenos que generan los datos, con el objetivo de testear diferentes hipótesis acerca de dichos parámetros.

**Data Mining**: Busca patrones, tendencias o relaciones que organicen o simplifiquen los datos con el objeto de hacerlos comprensibles.

**Psicología Cognitiva**: Propone comprender o al menos modelar los mecanismos subyacentes en el aprendizaje de los humanos.

**Teoría de la Ciencia**: Busca conformar teorías descriptivas o normativas acerca del proceso de generación y justificación del conocimiento.

# Tipos de aprendizaje

El aprendizaje se puede caracterizar en varias dimensiones. Una de ellas es entre aprendizaje **empírico** (se apoya en experiencia externa) y el **analítico** (se basa en la descripción del problema).

Otra distinción es entre aprendizaje **supervisado** (incluye casos donde el contexto y el resultado esperado son conocidos) y el **no supervisado**.

El modelo resultante puede ser **transparente** u **opaco** (caja blanca/negra).

Finalmente, tenemos el ML **simbólico** (cercano a lo formal/lógico) o el **numérico** (cercano al soft computing y al reconocimiento de patrones).

# ML numérico

**Clasificación**. Encontrar un modelo que permita predecir una categoría a partir de casos etiquetados

**Regresión**. A partir de muestras o mediciones, inducir una función que permita predecir valores o probabilidades.

**Clustering.** Encontrar descripciones compactas que cubra adecuadamente un conjunto de casos con propiedades conocidas.

**Reducción de dimensionalidad.** Encontrar una descripción más compacta del problema (lo cual facilita por ejemplo las tres tareas anteriores).

**Detección de atípicos.** Dererminar si un caso se comporta o no en la norma.

**Aprendizaje por refuerzo**. Se recompensa o castiga en función del éxito o fracaso en la tarea.

# ML Simbólico

**Sistemas Expertos**. Sistemas deductivos basados en reglas aprendidas previamente en forma asistida.

**Inducción**. Partir de casos particulares escogidos, para arribar a alguna generalización.

**Aprendizaje basado en casos**. Similar al anterior pero con algún grado de abstracción.

**Analogía**. Determinar una correspondencia o isomorfismo parcial entre dos representaciones.

**Descubrimiento**. Abducción no supervisada, usualmente en casos donde no se conocen propiedades o valores.

# Soft Computing

Tiene mayor foco en el proceso de **optimización** de la solución de un problema.

**Algoritmos genéticos y evolutivos**. Resuelven problemas de optimización con base en una metáfora de la evolución Darwiniana.

**Fuzzy logic**. Agregan cuantificación “posibilística” a los operadores de la lógica de predicados, con funciones entrenables.

**Metaheurísticas** (Ant Colony Optimization, Swarm Inteligence) también llamados modelos “biomorfos”, modelan la optimización a través de un proceso de emergencia colectiva.

# Data Mining

Tiene foco en descubrir **patrones** o tendencias en colecciones de datos masivas, con el objetivo de extraer y explotar el conocimiento descubierto.

Utiliza típicamente clustering para encontrar regularidades, regresión lineal, y detección de atípicos (similar a ML pero con un foco más estadístico).

Agrega además sumarización (o agregación), asociación, modelado de dependencias, etc.

Se diferencia además del ML numérico en su caracter automático o semi-automático.

# Áreas de aplicación del ML

| Business type/sector | Raw data examples | Business opportunities |
|---|---|---|
| **1. Bank, credit, and insurance** | Transaction history<br>Registration forms<br>External references such as the credit protection service<br>Micro and macro economic indices<br>Geographic and demographic data | Credit approval<br>Interest rates charges<br>Market analysis<br>Prediction of default<br>Fraud detection<br>Identifying new niches<br>Credit risk analysis |
| **2. Security** | Access history<br>Registration form<br>Texts of news and Web content | Pattern detection of physical or digital behaviors that offer any type of risk |
| **3. Health** | Medical records<br>Geographic and demographic data<br>Sequencing genomes | Predictive diagnosis (forecast)<br>Analysis of genetic data<br>Detection of diseases and treatments<br>Map of health based on historical data<br>Adverse effects of medication/treatments |
| **4. Oil, gas, and electricity** | Distributed sensor data | Optimization of production resources<br>Prediction/fault and found detection |

# Áreas de aplicación del ML

| Business type/sector | Raw data examples | Business opportunities |
|---|---|---|
| **5. Retail** | Transaction history<br>Registration form<br>Purchase path in physical and/or virtual stores<br>Geographic and demographic data<br>Advertising data<br>Customer complaints | Increasing sales by product mix optimization base on behavior patterns during purchase<br>Billing analysis (as-is, trends), the high volume of customers and transactions, credit profile by region<br>Increasing satisfaction/loyalty |
| **6. Production** | Data management systems/ERP production<br>Market data | Optimization of production over sales<br>Decreased time/amount of storage<br>Quality control |
| **7. Representative organizations** | Customer's registration form<br>Event data<br>Business process management and CRM systems | Suggestions of optimal combinations of company profiles, customers, and business leverage to suppliers<br>Synergy opportunities identification |
| **8. Marketing** | Micro and macroeconomic indices<br>Market research<br>Geographic and demographic data<br>Content generated by users<br>Data from competitors | Market segmentation<br>Optimizing the allocation of advertising resources<br>Finding niche markets<br>Performance brand/product<br>Identifying trends |

# Áreas de aplicación del ML

| Business type/sector | Raw data examples | Business opportunities |
|---|---|---|
| **9. Education** | Transcripts and frequencies<br>Geographic and demographic data | Personalization of education<br>Predictive analysis for school evasion |
| **10. Financial/ economic** | List of assets and their values<br>Transaction history<br>Micro and macroeconomics indexes | Identify the optimal value of buying complex assets with many analysis variables (vehicles, real estate, stocks, etc.)<br>Determining trends in asset values<br>Discovery of opportunities |
| **11. Logistic** | Data products<br>Routes and delivery points | Optimization of good flows<br>Inventory optimization |
| **12. E-commerce** | Customer registration<br>Transaction history<br>Users' generated content | Increase free users' conversion rate for paying users by detecting the heavier preferences of users |
| **13. Games, social networks, and platforms** | Access history<br>Registration of users<br>Geographic and demographic data | Increase free user conversion rate for paying users by detecting the behavior and preferences of users |
| **14. Recruitment** | Registration of prospects employees<br>Professional history, CV<br>Connections on social networks | The person's profile evaluation for a specific job role<br>Criteria for hiring, promotions, and dismissal<br>Better allocation of human resources |

# Etapas de una aplicación de ML

Una demarcación más o menos arbitraria de las tareas involucradas es:

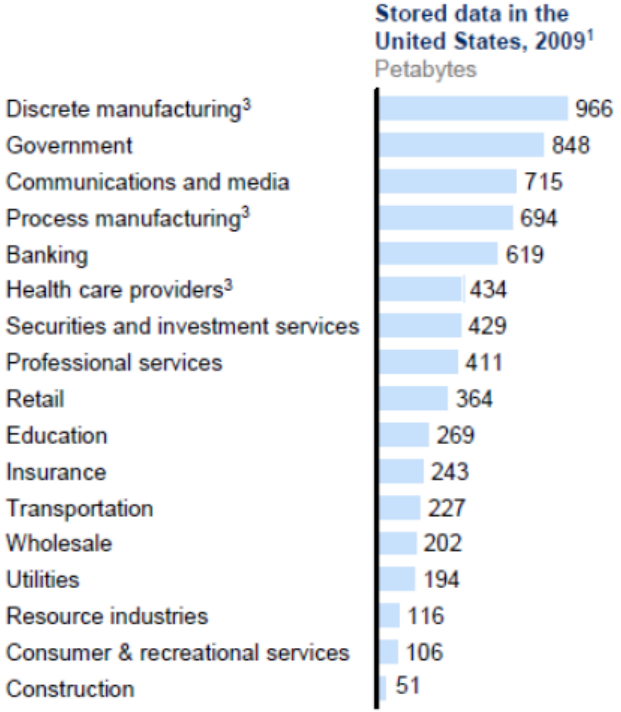
- Identificar la demanda e interpretar los requerimientos
- Identificar la fuente y procurar de los datos (data munging)
- Preparación de los datos (data wrangling)
- Análisis de datos (exploratorio, confirmatorio, modelado, implementación)
- Comunicación de los resultados (dashboards, visualización interactiva)
- Extras: documentación, seguridad, privacidad, legales





# Origen de los datos

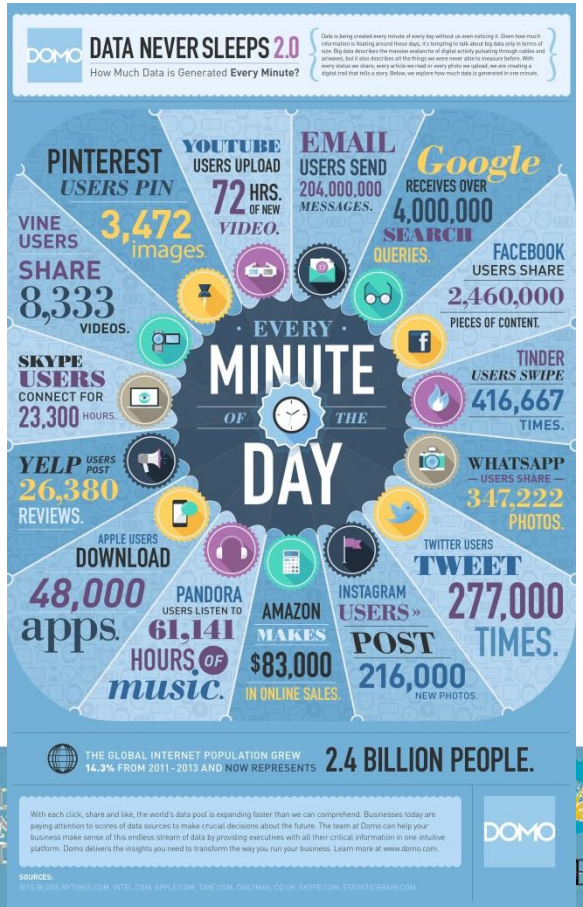
- Datos «legacy» vs. datos nativos digitales.
- Datos generados por humanos vs. generados por máquinas.
- Datos públicos vs. datos de pago (data markets).
- Datos corporativos, gubernamentales, científicos.
- Datos estructurados, semiestructurados, no estructurados.



# Origen de los datos

El streaming de datos se está convirtiendo en la fuente principal de datos

- Internet
- GPS
- IoT
- Cámaras de seguridad
- Smart meters
- Vehículos (aviones por ejemplo)
- ALMA, SHC
- etc. etc.



# ML y Big Data

## VELOCIDAD

El contenido digital a nivel mundial se **duplicará cada 18 meses.**

IDC


## VOLUMEN

En 2005, la humanidad creó 150 Exabytes de Información; en 2011 se crearon **1,200 Exabytes.**

The Economist

## VARIEDAD

**80% de los datos empresariales serán no estructurados** desde fuentes tanto tradicionales como no tradicionales.

Gartner

# Preparación de los datos (wrangling)

Inicialmente vamos a pensar que nuestro **dataset** está en un formato tabular. Cada fila es un **registro** (o dato u observación), cada columna es un **atributo** (o dimensión o variable), y cada celda es un **valor**.

Los atributos pueden ser nominales, binarios, ordinales, cuantitativos…

Valor

Atributo

Registro

| Name | Thread pitch (mm) | Minor diameter tolerance | Nominal diameter (mm) | Head shape | Price for 50 screws | Available t factory outlet? | Number in stock | Flat or Phillips head? |
|---|---|---|---|---|---|---|---|---|
| M4 | 0.7 | 4g | 4 | Pan | [illegible] | Yes | 276 | Flat |
| M5 | 0.8 | 4g | 5 | Round | $13.89 | Yes | 183 | Both |
| M6 | 1 | 5g | 6 | Button | $10.42 | Yes | 1043 | Flat |
| M8 | 1.25 | 5g | 8 | Pan | $11.98 | No | 298 | Phillips |
| M10 | 1.5 | 6g | 10 | Round | $16.74 | Yes | 488 | Phillips |
| M12 | 1.75 | 7g | 12 | Pan | $18.26 | No | 998 | Flat |
| M14 | 2 | 7g | 14 | Round | $21.19 | No | 235 | Phillips |
| M16 | 2 | 8g | 16 | Button | $23.57 | Yes | 292 | Both |
| M18 | 2.1 | 8g | 18 | Button | $25.87 | No | 664 | Both |
| M20 | 2.4 | 8g | 20 | Pan | $29.09 | Yes | 486 | Both |
| M24 | 2.55 | 9g | 24 | Round | $33.01 | Yes | 982 | Phillips |
| M28 | 2.7 | 10g | 28 | Button | $35.66 | No | 1067 | Phillips |
| M36 | 3.2 | 12g | 36 | Pan | $41.32 | No | 434 | Both |
| M50 | 4.5 | 15g | 50 | Pan | $44.72 | No | 740 | Flat |

# Preparación de los datos (wrangling)

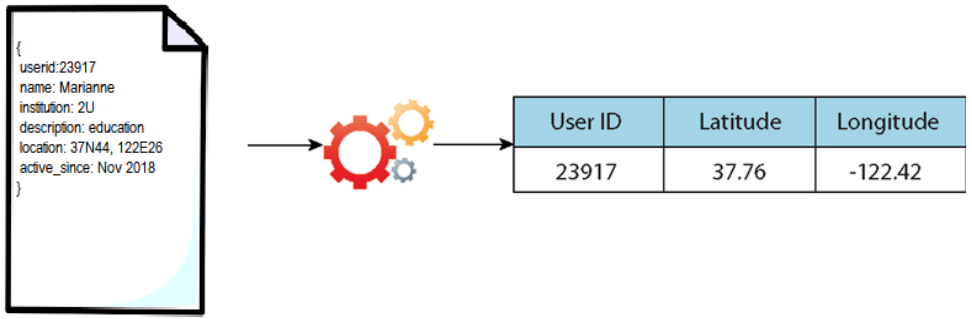
Usualmente para arribar a un dataset de esas características se sigue una serie de pasos: adquisición, filtrado, extraccion, validación, y agregación. Estos procesos están “relativamente” estandarizados y se evalúan de acuerdo a criterios de calidad de datos.

Salvo casos mínimos, conviene sistematizar y automatizar el flujo de trabajo a través de procesos de **gobernanza**, lo cual otorga trazabilidad, auditabilidad, y varias otras ventajas. Finalmente, agregar headers con metadata a los datasets para poder determinar los procesos involucrados.

# Preparación de los datos (wrangling)

Los datos no estructurados (video, audio, texto libre, etc.) son los más abundantes, pero para extraer información de ellos se necesitan "analíticos" que en general son ad-hoc, no escalan, no articulan.

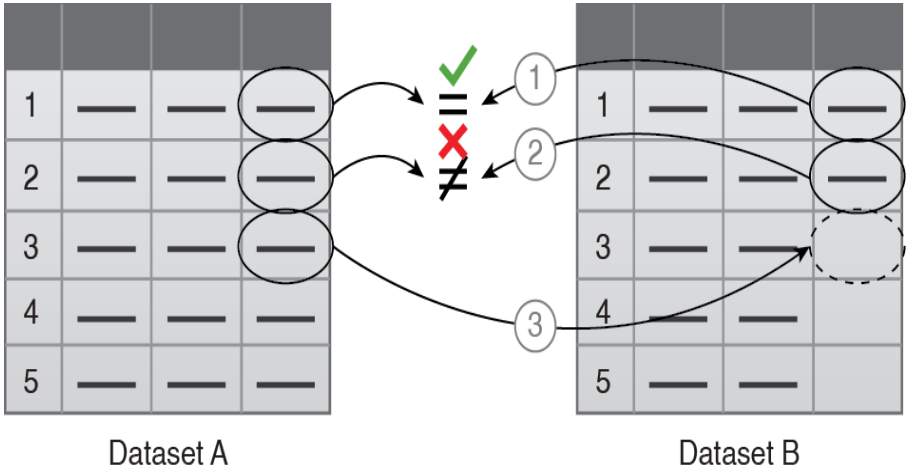
Los datos semiestructurados (JSON, XML, dataframes, "schema-less") son más fácilmente tratables, pero su estructura no se conoce estáticamente.




Los datos estructurados (.csv, SQL, tabulares) requieren menos preparación.

# Preparación de los datos (wrangling)

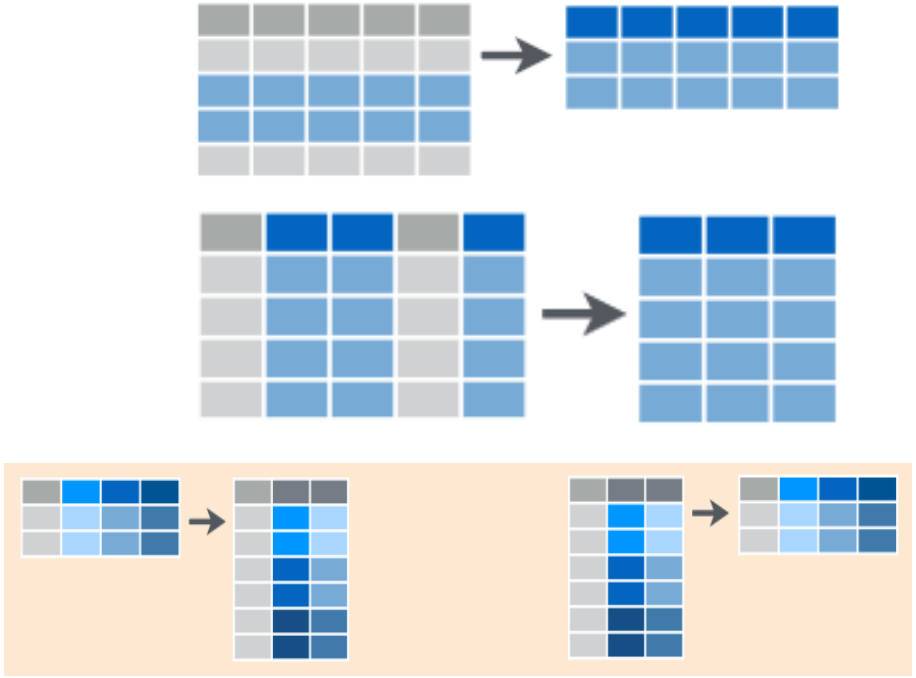
Las tareas de agregación e imputación son típicas de la conciliación de dos o más datasets. En estos casos existen varias maneras "correctas" posibles, dependiendo mucho del contexto. Por ello la necesidad de tener gobernanza clara, y metadata para realizar trazabilidad y "foldback" en casos de error.



# Preparación de los datos (wrangling)

Las tareas de filtrado, reducción de la dimensionalidad (por proyección u otras), o *reshaping* del dataset son típicas de la preparación previa al análisis, para conformar el dataset a las restricciones de la técnica de análisis que se utilice.

Esto depende asimismo del "ancho" y "alto" del dataset.



# Ejemplos de procurar y preparar datos

Ver los siguientes notebooks:

https://github.com/manlio99/Materia-de-aprendizaje/blob/master/4_DataWrangling/01_data_extraction_sat_esp.ipynb

https://github.com/manlio99/Materia-de-aprendizaje/blob/master/4_DataWrangling/02_data_extraction_texas_death_row_executions_esp.ipynb

https://github.com/manlio99/Materia-de-aprendizaje/blob/master/4_DataWrangling/03_data_extraction_amazon_esp.ipynb

https://github.com/manlio99/Materia-de-aprendizaje/blob/master/4_DataWrangling/04_data_extraction_nasa_esp.ipynb

# Ejercicio para pensar

Sean los siguientes cuatro datasets (conocidos como el cuarteto de Anscombe).

| I | |
|---|---|
| x | y |
| 10.0 | 8.04 |
| 8.0 | 6.95 |
| 13.0 | 7.58 |
| 9.0 | 8.81 |
| 11.0 | 8.33 |
| 14.0 | 9.96 |
| 6.0 | 7.24 |
| 4.0 | 4.26 |
| 12.0 | 10.84 |
| 7.0 | 4.82 |
| 5.0 | 5.68 |

| II | |
|---|---|
| x | y |
| 10.0 | 9.14 |
| 8.0 | 8.14 |
| 13.0 | 8.74 |
| 9.0 | 8.77 |
| 11.0 | 9.26 |
| 14.0 | 8.10 |
| 6.0 | 6.13 |
| 4.0 | 3.10 |
| 12.0 | 9.13 |
| 7.0 | 7.26 |
| 5.0 | 4.74 |

| III | |
|---|---|
| x | y |
| 10.0 | 7.46 |
| 8.0 | 6.77 |
| 13.0 | 12.74 |
| 9.0 | 7.11 |
| 11.0 | 7.81 |
| 14.0 | 8.84 |
| 6.0 | 6.08 |
| 4.0 | 5.39 |
| 12.0 | 8.15 |
| 7.0 | 6.42 |
| 5.0 | 5.73 |

| IV | |
|---|---|
| x | y |
| 8.0 | 6.58 |
| 8.0 | 5.76 |
| 8.0 | 7.71 |
| 8.0 | 8.84 |
| 8.0 | 8.47 |
| 8.0 | 7.04 |
| 8.0 | 5.25 |
| 19.0 | 12.50 |
| 8.0 | 5.56 |
| 8.0 | 7.91 |
| 8.0 | 6.89 |

# Ejercicio para pensar

Supongamos que tenemos cuatro nuevos datos ([16,8]; [8,6]; [16,10] y [16,6]) y sabemos que cada uno corresponde a un único dataset. Sin embargo, los parámetros estadísticos de los cuatro datasets son idénticos (ver debajo). Cómo determinar entonces a qué dataset corresponde cada punto?

| Propiedad | Valor |
|---|---|
| Media de x | 9 |
| Variancia de x | 11 |
| Media de y | 7.50 |
| Variancia de y | 4.125 |
| Correlación entre x e y | 0.816 |
| Linea de regresión | y = 3.00 + 0.500x |
| Coeficiente de determinación $R^2$ | 0.67 |

# Bibliografía


- Raschka S. Python machine learning: unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics.Packt Publishing 2015.
- Harrington P. Machine learning in action. N.Y., Manning Publications 2012.
- Wes McKinney. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.), O'Reilly 2018.
- Zhang, A. (2018). Data analytics: Practical guide to leveraging the power of algorithms, data science, data mining, statistics, Big Data, and predictive analysis to improve business, work, and life. Kindle Edition.