

Estadísticas para Data Science

Ing. Julio Paciello

juliopaciello@cds.com.py

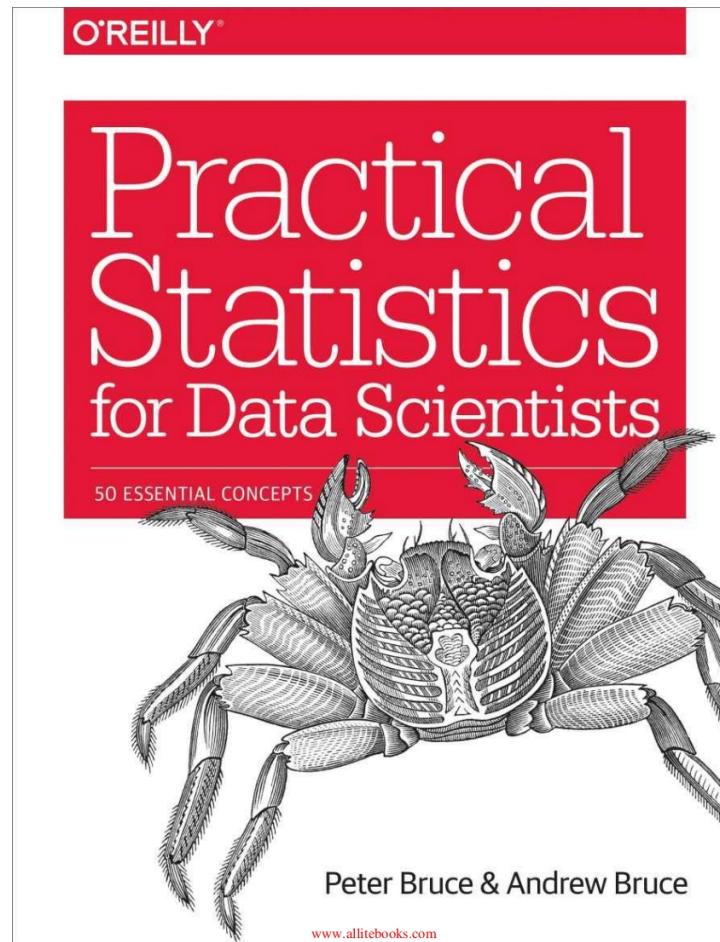
Contenido

Data Science

- Terminología y conceptos
- Base conceptual estadística
- Tipos de Análisis de Datos
- Análisis Exploratorio (EDA)
 - 1 variable
 - Varias variables
- Correlación de variables
- Inferencia Estadística

Bibliografía

Bruce, P. & Bruce, A. Practical Statistics for Data Scientists



Mi experiencia previa



infobae

LUNES 24 DE SEPTIEMBRE DE 2018
PRECIO UNICO \$100 - AÑO D 11 - 35

SUDAMERIS

Seguridad Seguros

Grupo Cavallaro

Automaq
Automóviles y Maquinaria Agrícola

Tel.: 021 379 9000

RÍOS
REPUESTOS RG S.A.
Tel.: 021 511 296

#EDITORIAL
Hay un compromiso asumido por el Paraguay de aplicar estrictas normas de protección ambiental y de manejo sostenible de los recursos naturales y de las actividades económicas. El GATF señala claramente que el desarrollo sostenible es una estrategia en esta lucha, así como el combate al cambio climático.

SEGUIMIENTO AL PLAN NACIONAL DE DESARROLLO SOSTENIBLE

EL RETRASO DE LOS PAGOS EN LAS CONTRATACIONES PÚBLICAS SIGNIFICA UN COSTO DE ALGUNOS US\$ 20 MILLONES AL AÑO.

PRINCIPAL PROBLEMA SE DA EN LA CARGA DE LAS FACTURAS. PÉRDIDA SIGNIFICA EL DÉBITO DE QUE COSTÓ LA COSTA NERA. NUEVA ASUNCIÓN, DATOS REVELAN QUE INEFICIENCIA CRECIÓ EN LOS ÚLTIMOS 7 AÑOS.

> PÁGS. 6 Y 7

5días
PASIÓN POR LOS NEGOCIOS

COBERTURA ESPECIAL
Marito se reúne con tenedores de bonos > PÁG. 10

INEFICIENCIA
pública cuesta más de US\$ 140 millones

SEGÚN DATOS DEL CENTRO DE DESARROLLO SOSTENIBLE EL RETRASO DE LOS PAGOS EN LAS CONTRATACIONES PÚBLICAS SIGNIFICA UN COSTO DE ALGUNOS US\$ 20 MILLONES AL AÑO.

PRINCIPAL PROBLEMA SE DA EN LA CARGA DE LAS FACTURAS. PÉRDIDA SIGNIFICA EL DÉBITO DE QUE COSTÓ LA COSTA NERA. NUEVA ASUNCIÓN, DATOS REVELAN QUE INEFICIENCIA CRECIÓ EN LOS ÚLTIMOS 7 AÑOS.

> PÁGS. 6 Y 7

AGRONEGOCIOS GANADEROS PARAGUAYOS BUSCAN SEDUCIR AL MERCADO AMERICANO > PÁG. 21

COMERCIO ESPERAN QUE CRISIS ARGENTINA SE DISPARE CON ESPALDARAZ DEL FMI > PÁG. 4

DÓLAR 5.790 5.830 | PESO 147 150 | REAL 1.375 1.435 | EURO 0.580 0.600 | BOVESPA 79.444 | DOW JONES 26.743

LA PRIMAVERA ESTÁ EN EL AIRE.
Y USTED EN EL NUEVO BMW X3.

Disfrute la primavera más divertida del año con su nuevo BMW X3 desde \$1.900 USD.
Las unidades son limitadas, complete el deseo o v por nuevos descuentos!

BMW Gallery | Av. Sacramento 347 | +595 21 689 3000.
Sede Hermandad 1100 | Parque Industrial de los Ángeles | +595 81 379 750.

Perfecta Automotriz S.A.

BMW X3



TETÃ VIRU MOHENDAPY
MOTENONDEHA
MINISTERIO DE HACIENDA

Personal



VUE
VENTANILLA ÚNICA DE EXPORTACIÓN

Materiales y Evaluación

- Repositorio GIT:
 - <https://github.com/cdsparaguay/cursoml>
- Presentaciones
- Prácticas:
 - Open Contracting Data Standard (OCDS)
 - EDA en R
- Bibliografía

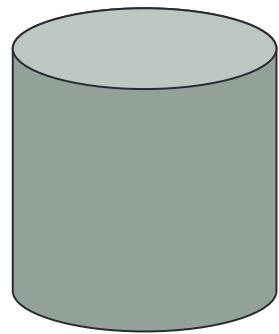
Terminología y conceptos

Business Intelligence

- Infraestructura tecnológica para obtener la máxima información de los datos disponibles para la mejora continua de los procesos de negocio
- Sistemas de BI
 - OLAP (Online Analytical Processing)
 - CRM (Customer Relationship Management)
 - GIS (Geographic Information System)
 - **KDD (Knowledge Discovery in Databases)**
 - ...

Business Intelligence

Aggregate Data



Present Data



Enrich Data



Inform a Decision



Database, Data Mart, Data Warehouse, ETL Tools, Integration Tools

Reporting Tools, Dashboards, Static Reports, Mobile Reporting, OLAP Cubes

Add Context to Create Information, Descriptive Statistics, Benchmarks, Variance to Plan or LY

Decisions are Fact-based and Data-driven

Amazon.com and NetFlix

Collaborative Filtering

Intentan predecir otros ítems que un cliente desea comprar en base a lo que hay en sus shopping cart y wish lists y el comportamiento de compra de otros clientes

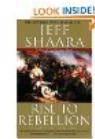
Customers Who Bought This Item Also Bought



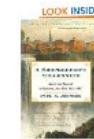
The Last Full Measure by Jeff Shaara
★★★★★ (149)
\$7.99



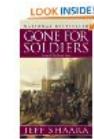
Gods and Generals by Jeff Shaara
★★★★★ (248)
\$7.99



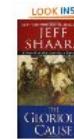
Rise to Rebellion: A Novel of the American Revolution by Jeff Shaara
★★★★★ (162)
\$10.85



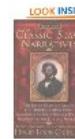
A Shopkeeper's Millennium: Society and Rev... by Paul E. Johnson
★★★★★ (9)
\$11.20



Gone For Soldiers by Jeff Shaara
★★★★★ (108)
\$7.99



The Glorious Cause by Jeff Shaara
★★★★★ (84)
\$7.99



The Classic Slave Narratives-paperback by Henry Louis Gates
★★★★★ (11)
\$7.95

House of cards

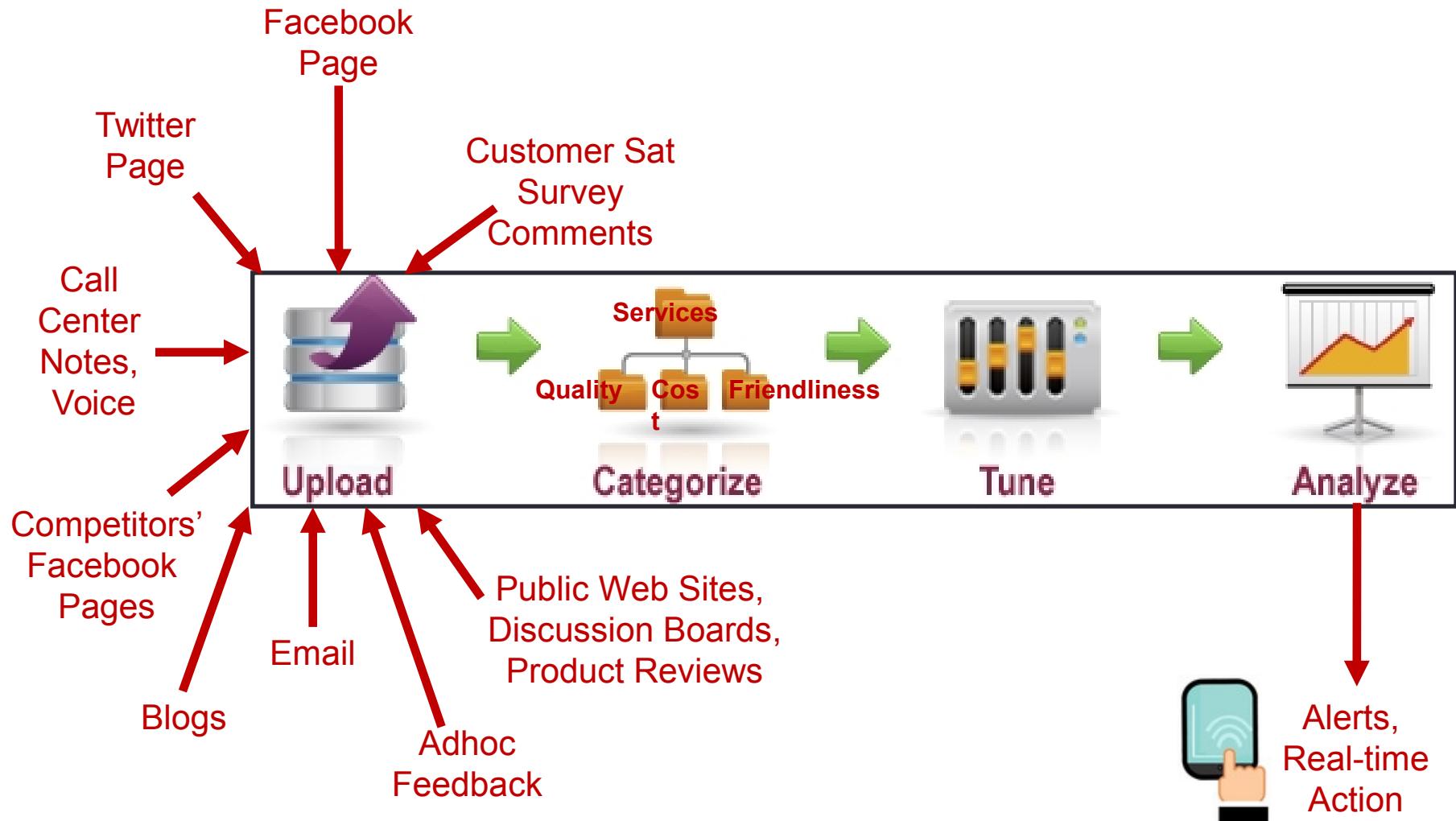


Referencias:

<https://www.cio.com/article/3207670/big-data/how-netflix-built-a-house-of-cards-with-big-data.html>

<https://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html>

Unstructured Text Processing

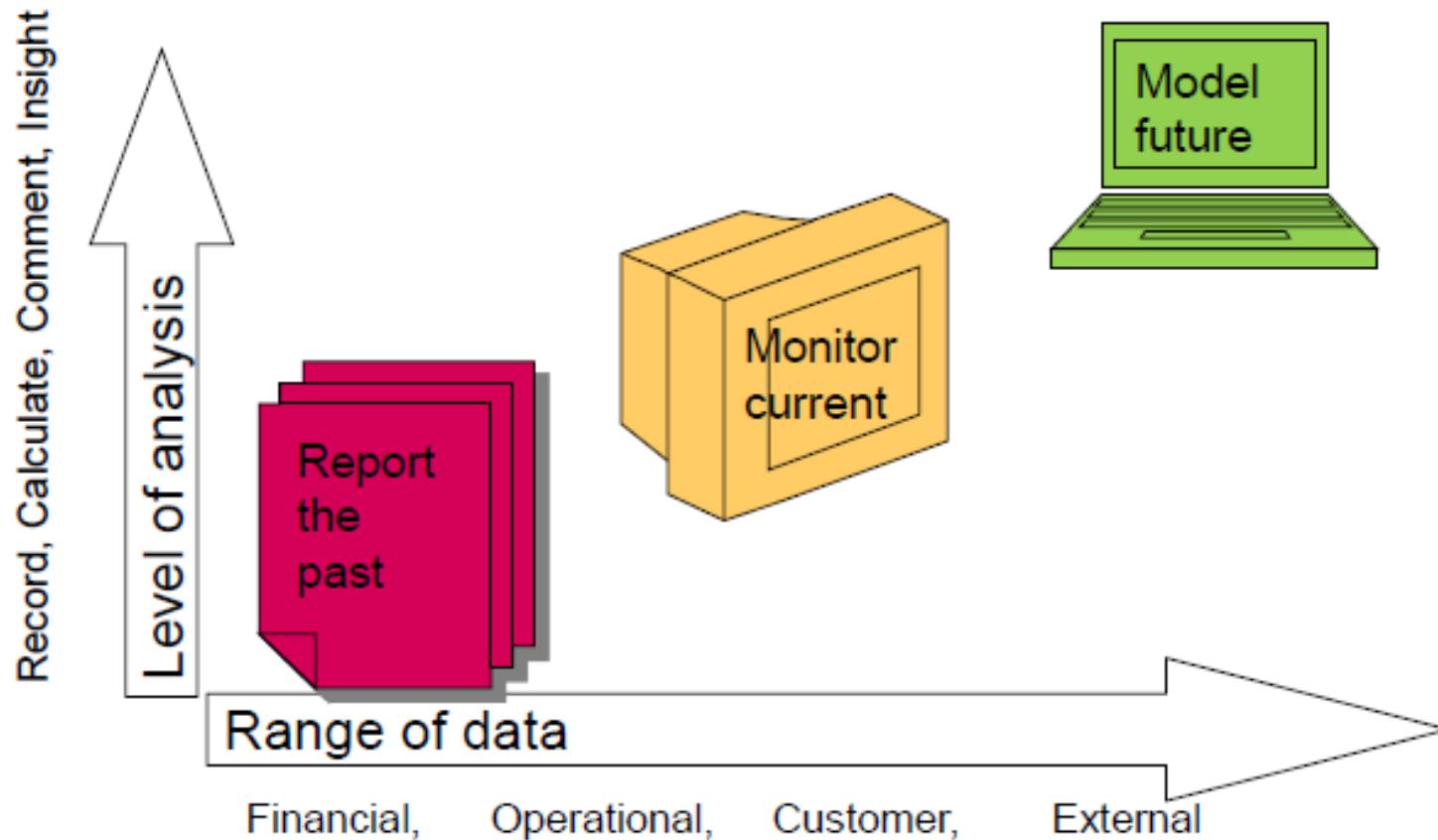


Unstructured Text Processing



<https://www.predictiveanalyticstoday.com/lexalytics/>

Niveles de análisis (madurez)

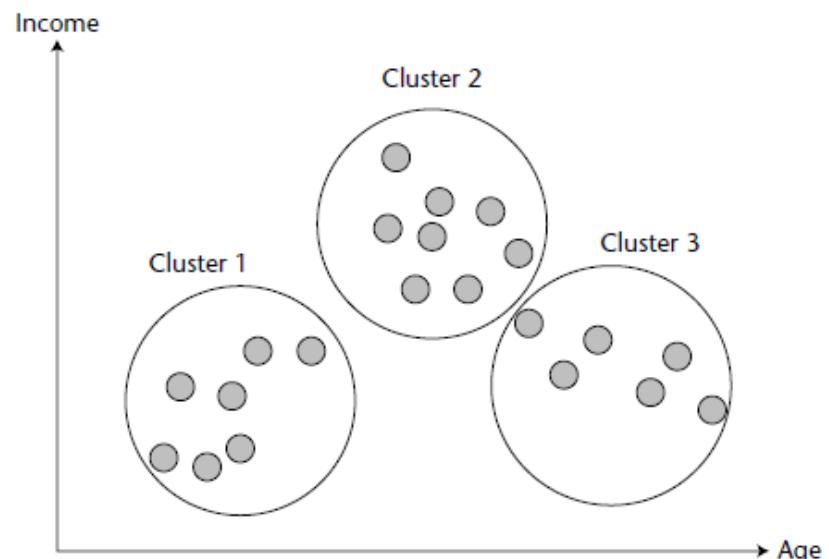
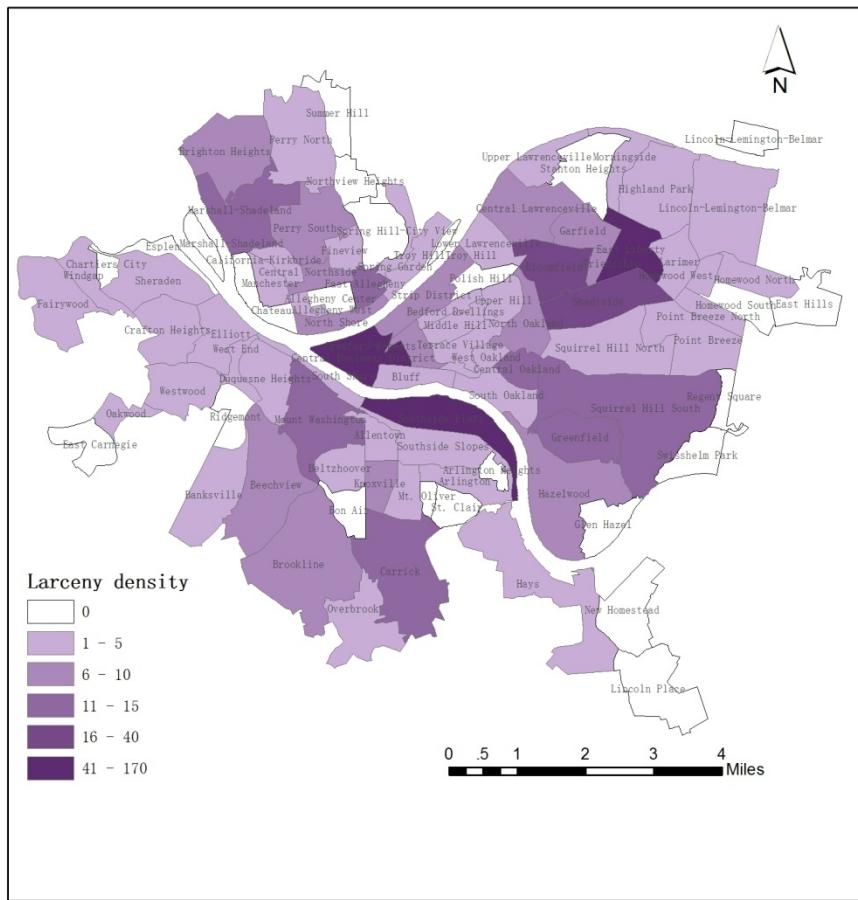


Data Mining

- **Estadísticas:** regresión, análisis multivariable, análisis clúster, clasificación
- **Técnicas de inteligencia artificial:** redes neuronales, algoritmos predictivos

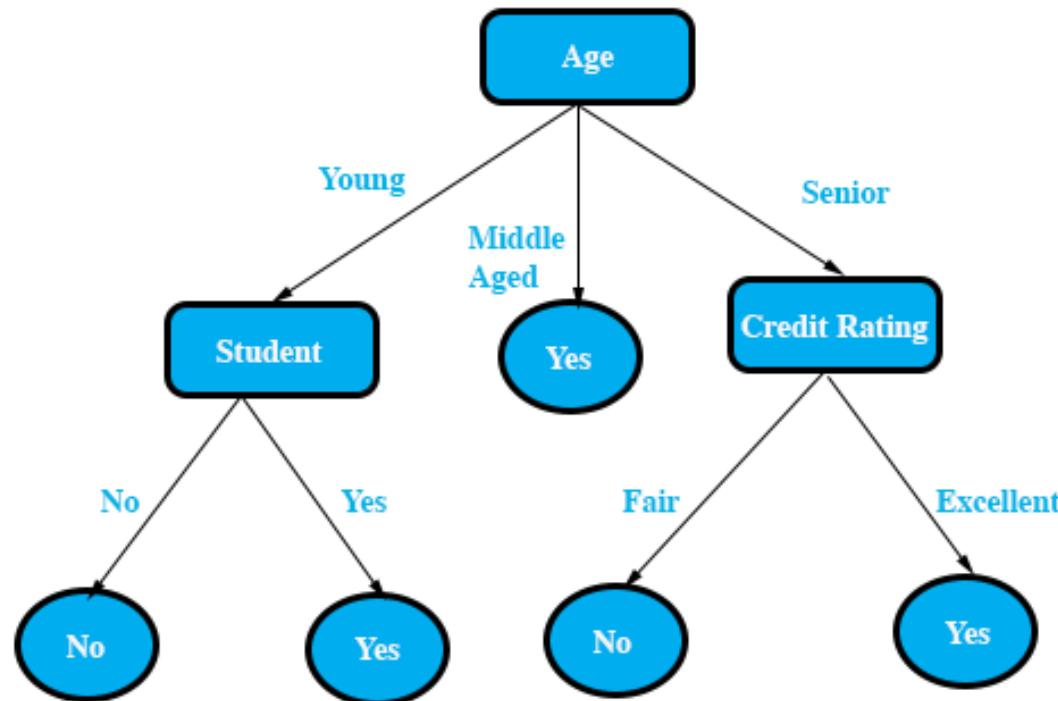
Data Mining

- Clustering



Data Mining

- Decision Trees
 - Ej: Aprobación de crédito



Data Mining

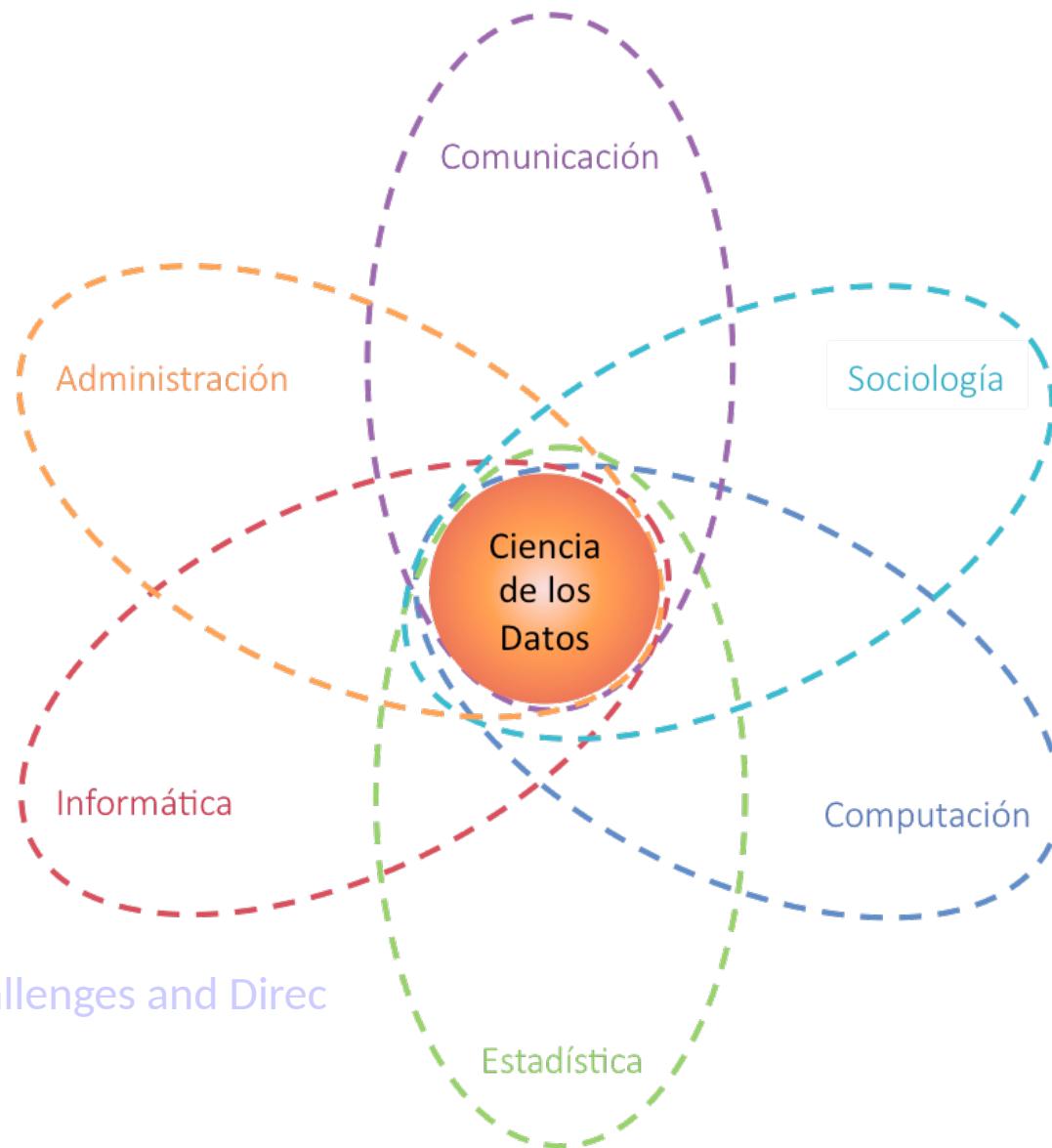
- Red Neuronal



PowerAI

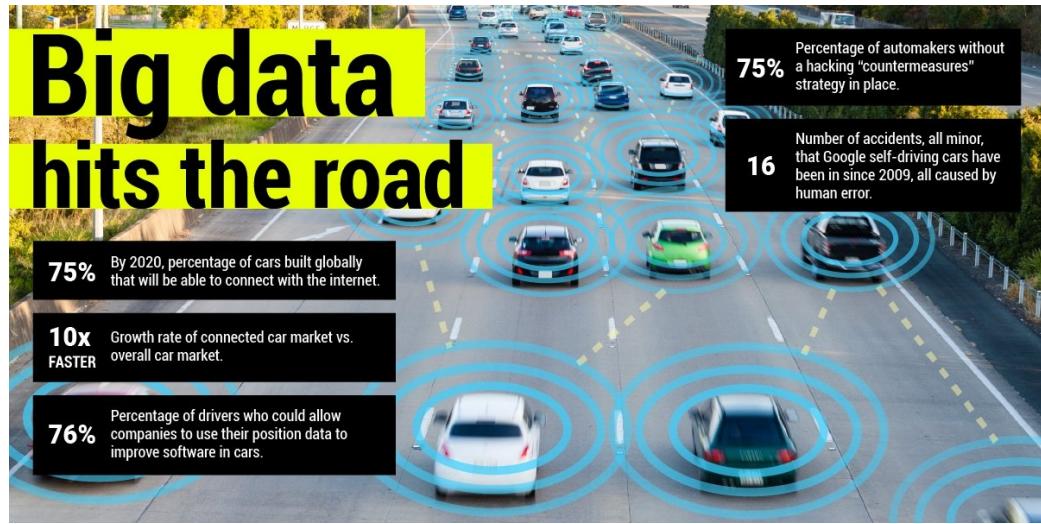
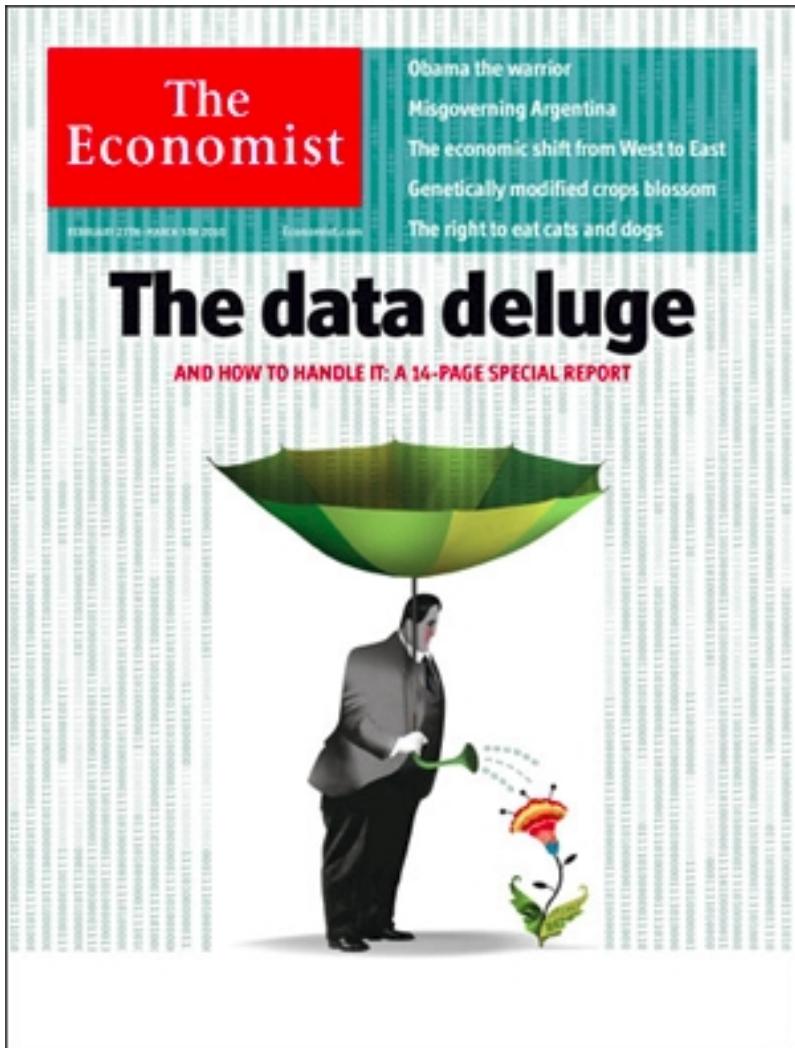
<https://www.youtube.com/watch?v=0F5w6q0ZpBI>

Qué es Data Science



Data Science: Challenges and Directions

Motivación



El big data es el motor de la innovación actual (self-driving cars, real-time object classifiers, voice assistants, chatbots)

Big Data Hits The Road

Motivación

La demanda de científicos de datos ha crecido exponencialmente desde el 2011 ([The number one job in America](#))



Datos (el nuevo petróleo)



The world's most valuable resource is no longer oil, but data

Datos

Los **datos** son valores pertenecientes a **conjuntos de ítems** y representados a través de **variables cualitativas o cuantitativas**

Conjunto de items: Conjunto de objetos de interés, a veces llamado población

Variables: Medida o característica de un ítem

Cualitativo: País de origen, sexo, religión

Cuantitativo: Altura, peso, edad

Datos

ID	Latitud	Longitud	Velocidad	Fecha
Juan	-25.291455	-57.586020	27 km/h	Lunes, 20/Marzo/17 08:05:25
Juan	-25.294947	-57.578006	31 km/h	Lunes, 20/Marzo/17 08:35:25
Juan	-25.291959	-57.575141	35 km/h	Lunes, 20/Marzo/17 09:10:15
Juan	-25.287312	-57.572287	28 km/h	Lunes, 20/Marzo/17 09:15:20
Claudia	-25.286352	-57.587393	38 km/h	Lunes, 20/Marzo/17 08:25:12
Claudia	-25.285886	-57.574529	32 km/h	Lunes, 20/Marzo/17 08:36:20
Claudia	-25.285964	-57.571922	42 km/h	Lunes, 20/Marzo/17 09:02:30
Claudia	-25.285188	-57.571729	40 km/h	Lunes, 20/Marzo/17 09:07:35

Información

Tramo	ID	Vel. Prom.	Día	Hora
Mcal Lopez	Juan	29 km/h	Lunes	8-10 am
San Martín	Juan	32 km/h	Lunes	8-10 am
España	Claudia	35 km/h	Lunes	8-10 am
San Martín	Claudia	41 km/h	Lunes	8-10 am

Datos procesados con alguna semántica

Conocimiento

MAPA DE TRÁFICO

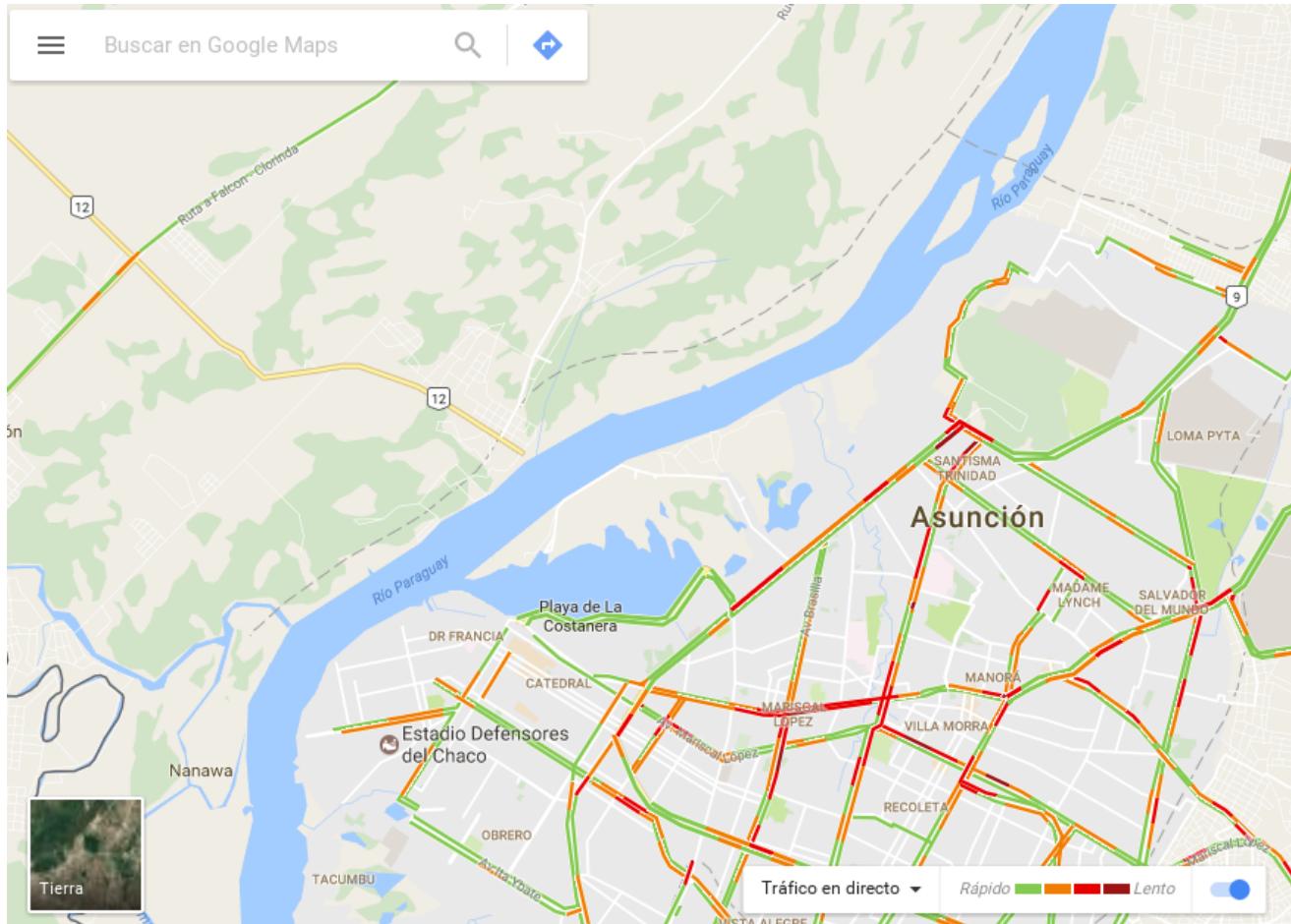
0-30 : rojo

31-50: naranja

>50 : verde

DIA: Lunes

HORA: 8 a 10 am



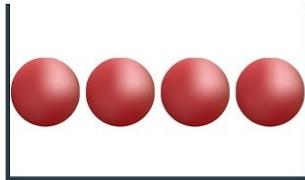
Decido por donde ir!

Base conceptual estadística

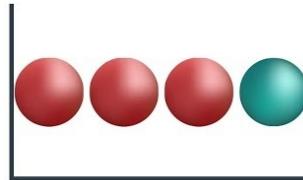
Teoría de la información

Claude Shannon

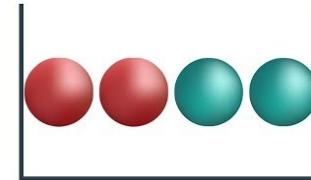
Entropy



Low



Medium



High



Entropía de la información

- *Entropía de Shannon*: cantidad de información promedio que contienen los símbolos utilizados. Aquellos símbolos con menor probabilidad de aparición aportan más información. Ej: en un texto en español las palabras “que” “el” que tienen alta probabilidad de ocurrencia y por ende aportan baja información, si las eliminamos igual podría ser entendible el texto.
- En un lenguaje binario= $\{0,1\}$, cada valor posible posee un $1/2$ (50%) de probabilidad de ocurrencia, aportan la misma cantidad de información.

Entropía de la información

$$c_i = \log_2(k) = \log_2[1/(1/k)] = \log_2(1/p) = \underbrace{\log_2(1)}_{=0} - \log_2(p) = -\log_2(p)$$

$$H = -p_1 \log_2(p_1) - p_2 \log_2(p_2) - \dots - p_k \log_2(p_k) = - \sum_{i=1}^k p_i \log_2(p_i)$$

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

Es la suma de la cantidad de información aportada por cada símbolo ponderada por su probabilidad de ocurrencia

k : cantidad símbolos distintos

C_i : cantidad de información (bits) para representar el símbolo i

P_i : probabilidad de ocurrencia del símbolo i

Ejemplo de Entropía

- Asumamos una variable X con 3 valores posibles ALTO, MEDIO, BAJO con probabilidades de ocurrencia de 1/3 (33%), 1/2 (50%) y 1/6 (17%) respectivamente, la entropía H de la variable sería:

$$H(X) = \frac{1}{3} \log_2(3) + \frac{1}{2} \log_2(2) + \frac{1}{6} \log_2(6) = 1,46$$

Ejemplo de Entropía

- *Valor Máximo de H*, distribución uniforme, 1/3 (33%), 1/3 (33%), 1/3 (33%):
$$H(X) = 1/3 \log_2(3) + 1/3 \log_2(3) + 1/3 \log_2(3) = 1,58$$
- *Valor Mínimo de H*, distribución sesgada, 0%, 100%, 0%:
$$H(X) = 0 + 1 \log_2(1) + 0 = 0$$

Medidas de posición

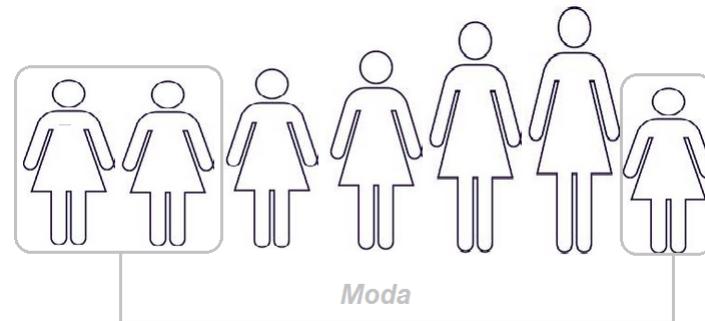
- Determinan una posición o ubicación dentro la distribución de los datos:
 - Tendencia central
 - No centrales
- Caracterizar la distribución de datos mediante un valor o rango representativo

Medidas de tendencia central

- Media aritmética ($X = \{x_1, x_2, \dots, x_n\}$)

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Moda ($X = \{x_1, x_2, \dots, x_n\}$)

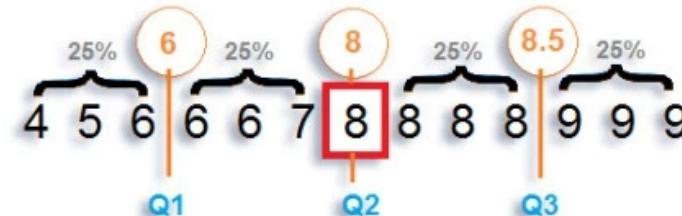


- Mediana ($X = \{x_1, x_2, \dots, x_n\}$)

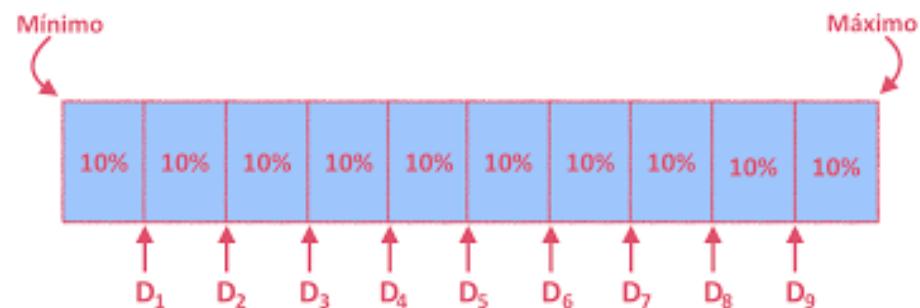


Medidas de posición no centrales

- Cuartiles ($X = \{x_1, x_2, \dots, x_n\}$)

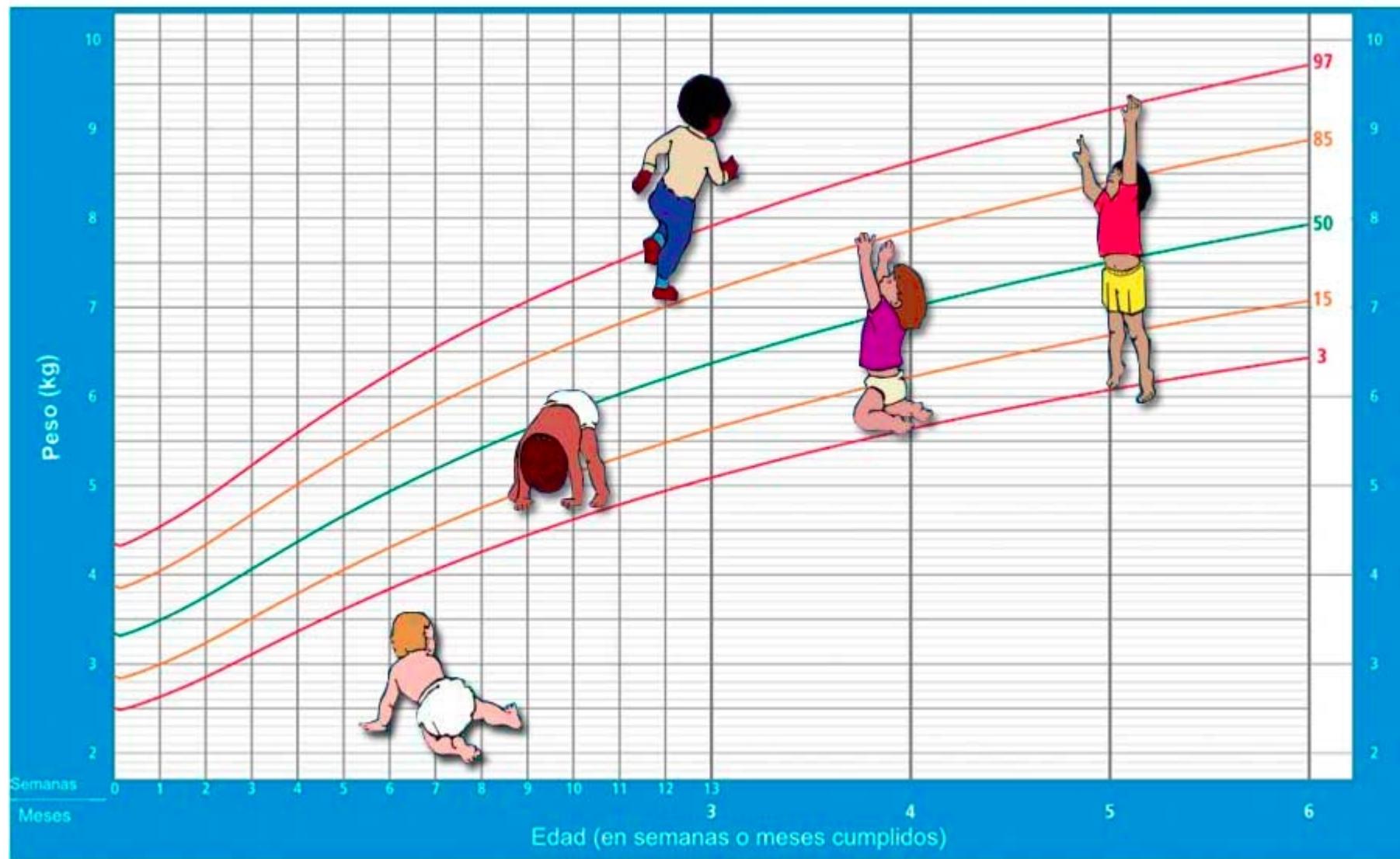


- Deciles ($X = \{x_1, x_2, \dots, x_n\}$)

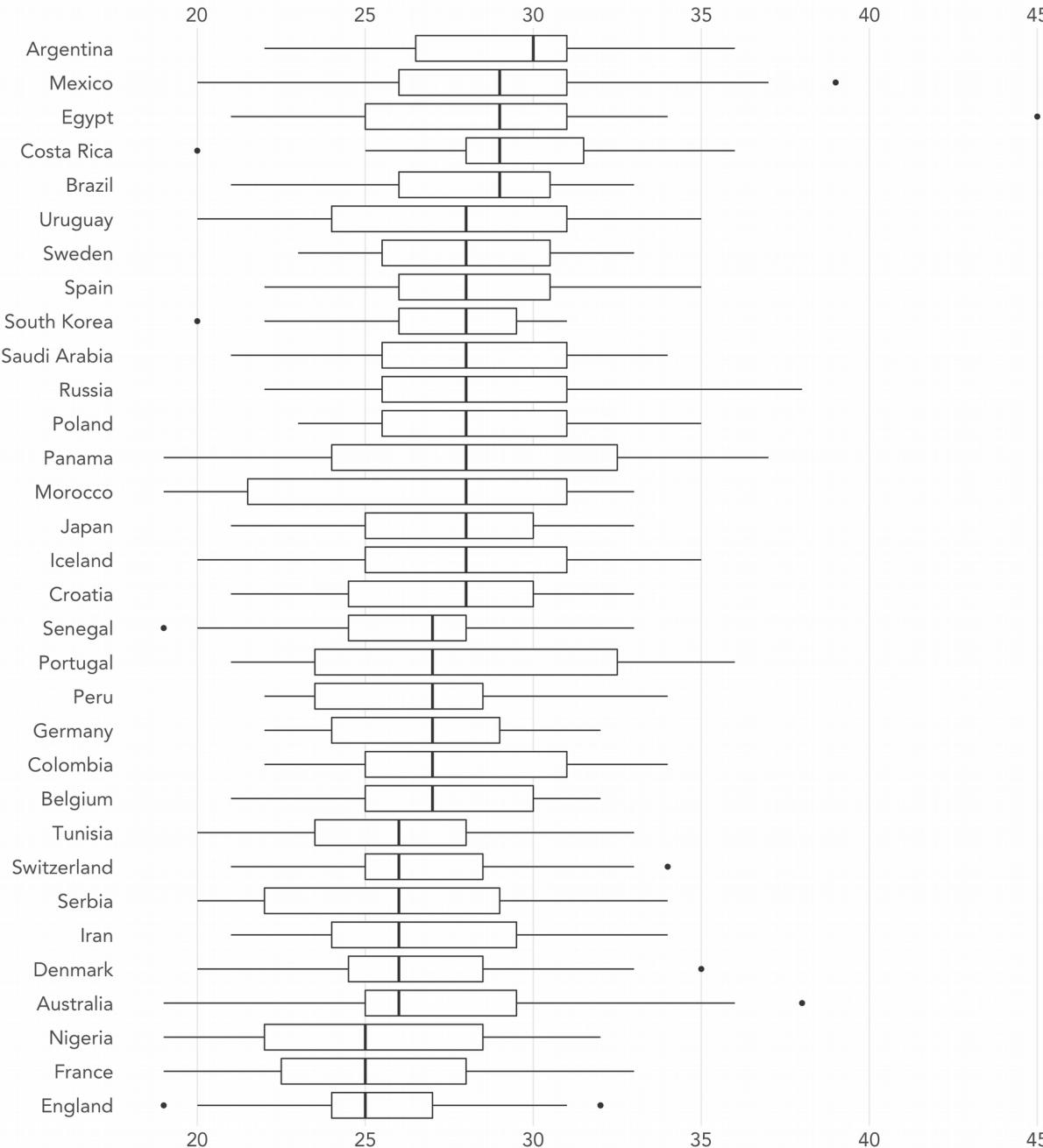


- Percentiles ($X = \{x_1, x_2, \dots, x_n\}$)

Medidas de posición no centrales



2018 Russia World Cup: Players' Age Distribution



2018 Russia
World Cup pl
ayers by Age

Medidas de dispersión

- Varianza ($X = \{x_1, x_2, \dots, x_n\}$)

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$$

- Desviación estándar ($X = \{x_1, x_2, \dots, x_n\}$)

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Covarianza (XY , con $X = \{x_1, x_2, \dots, x_n\}$ e $Y = \{y_1, y_2, \dots, y_n\}$)

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Medidas de dispersión

- **Correlación (XY, con X = {x₁,x₂,...,x_n} e Y = {y₁,y₂,...,y_n})**

En estadística una de las técnicas más comunes para el estudio de asociaciones entre variables es la correlación

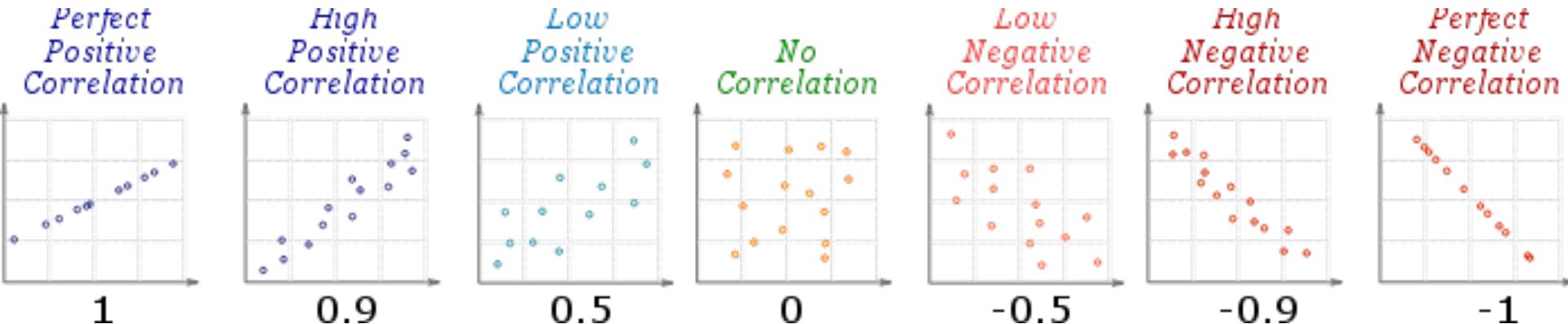
La correlación mide que tanto dos variables se encuentran linealmente relacionadas. A diferencia de la covarianza que determina simplemente si dos variables cambian a la par, la correlación mide la fuerza de la relación entre ellas

$$r = \text{Cov}(x, y) / S_x * S_y$$

Medidas de dispersión

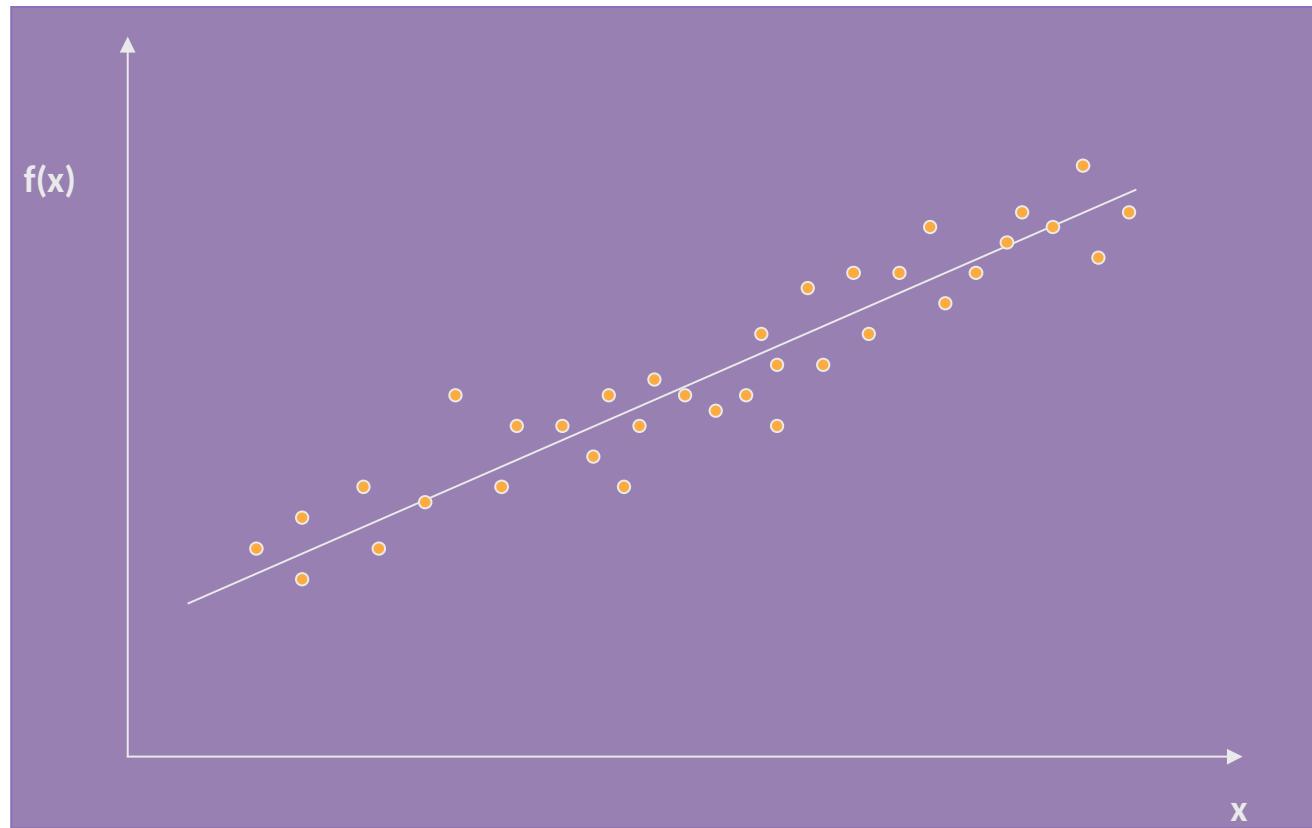
- Correlación (XY , con $X = \{x_1, x_2, \dots, x_n\}$ e $Y = \{y_1, y_2, \dots, y_n\}$)

El resultado de una análisis de correlación es un coeficiente (r o Person's r) entre -1 y 1. Los extremos indican correlación perfecta entre las variables. Una correlación negativa indica que las variables cambian en sentido contrario, mientras que una correlación positiva demuestra que ambas variables cambian en la misma dirección



Regresión Lineal

- **Tendencias / Regresión:** consiste en adquirir una función que mapee un elemento de dato a una variable de predicción de valor real, creando así un modelo predictivo



Tipos de variables

Según su influencia:

- **Variables Independientes (x)**: sus valores no están en dependencia de otras variables y creemos influyen en el fenómeno estudiado. Ej: clima, precios de canasta básica
- **Variables dependientes (y)**: sus valores están determinados por otras variables generalmente independientes. Ej: número de casos de dengue, IPC,

Tipos de variables

Según la Parametrización:

- **Variable Input:** Típicamente las variables independientes que influyen o creemos influyen en el modelo
- **Variable Target:** La variable de estudio que es influida por las variables independientes

Numerical Scoring

- Coeficiente de correlación de Pearson (ρ)

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- Coeficiente R^2

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}$$

- Mean Square Error (MSE)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- Root Mean Square Error (RMSE o RMSD)

$$\text{RMSD}(\hat{\theta}) = \sqrt{\text{MSE}(\hat{\theta})}$$

- Mean Absolute Error (MAE)

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

Referencias

- Bruce, P., Bruce A. Practical Statistics for Data Scientists: 50 Essential Concepts, 1st Ed, O'Really Media, 2017
- Larose, D. Discovering Knowledge in Data: An introduction to Data Mining. 1st Ed, Wiley. 2005
- Han, J., Kamber, M. Data Mining: Concepts and Techniques. 2nd Ed, Morgan Kaufmann. 2006