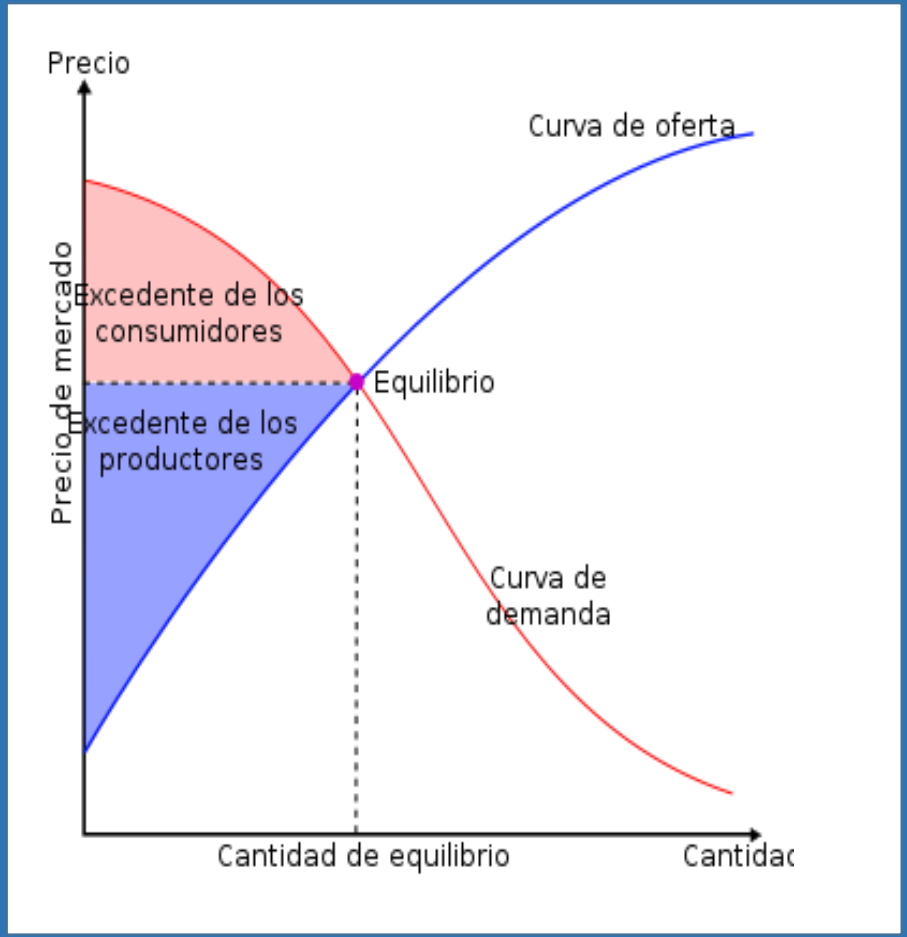
El Científico de Datos


ANALYTICS
INFORMATION
SHARED
TIBCO
Networks
trend
Science
Asset
capture
volume
Large
size
business
INTERNET
BIG DATA
cloud
SOFTWARE
CLUSTER
management
NAS
Is your Job sexy?
visualization
needed
storage
data
Databases
amount
quantity
complex
research
ANALYZING the ANALYZERS


Precio
Curva de oferta
Excedente de los consumidores
Equilibrio
Precio de mercado
Excedente de los productores
Curva de demanda
Cantidad de equilibrio
Cantidad

Científico de datos: profesión con un futuro BIG

por Óscar Méndez

Madrid, España | 14 de octubre del 2015 | big data, Big Data Spain, empleo | 1 comentario

Actualmente estamos en la tercera generación de IT. Resumiendo mucho, la primera fueron los grandes servidores hosts, la segunda los sistemas cliente/servidor con clusters de dos o pocas

TIBCO®

GERENCIA

EXPOENTIC.16 26 de OCTUBRE 2016

INICIO | NOTICIAS | EVENTOS | ARCHIVO | CLASIFICADOS | SUSCRIPCIONES | AVISADORES | PERFIL DEL MEDIO | CONTACTO | Búsqueda

Miércoles 19 de Octubre de 2016 • Dólar= $667,76 • UF=$26.241,21 • UTM=$45.999

Recomendar | Twittear | Comentar

DATA SCIENTIST O CIENTÍFICO DE DATOS

Cómo ganar dinero con las habilidades adecuadas

El volumen de datos crece y también la necesidad de quien analice toda esa información. El "McKinsey Global Institute" estima que unas 500 mil personas serán empleadas como científicos de datos para fines de esta década.

cibersur

DOSSIER | ENCUENTROS | CIBERSUR TV

Portada | Empresa | Internet | Comunicaciones | I+D+i | Software | Hardware | i-Administ

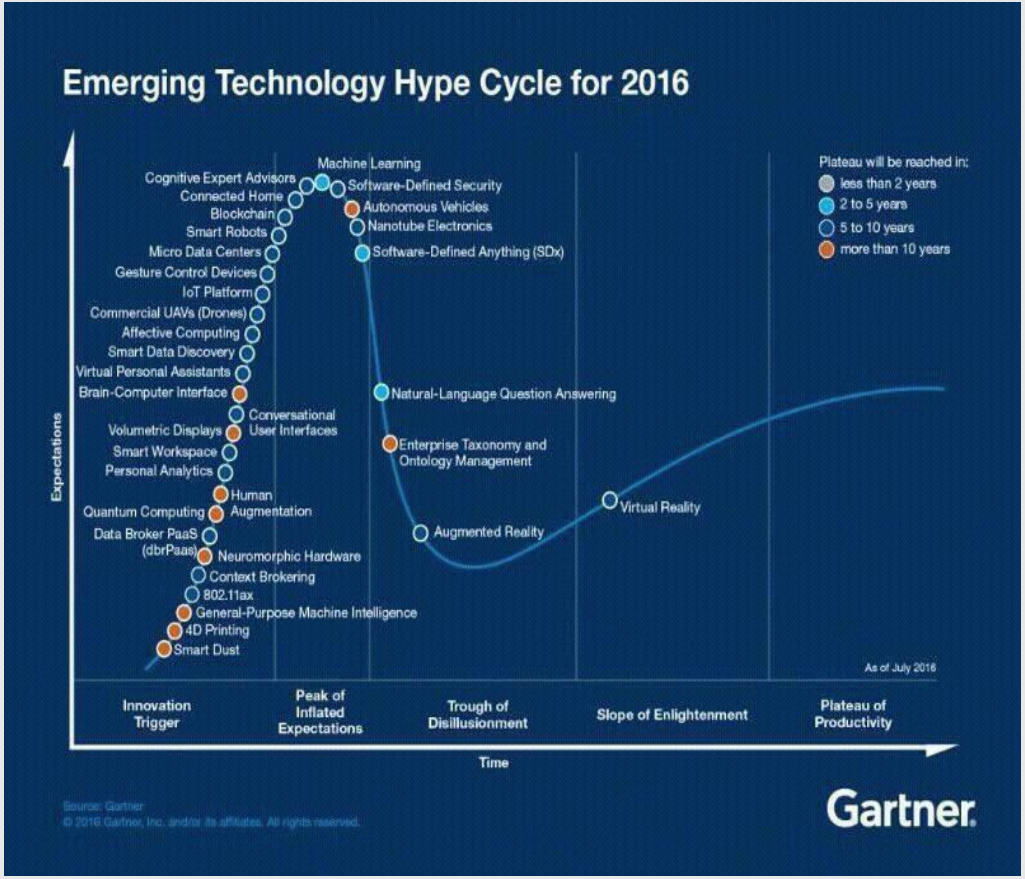
cibersur.com ▸ e-Sociedad

E-RRHH

El científico de datos, entre los de mayor demanda de empleo

Nadie duda de la creciente importancia de los científicos de datos (data scientists) en las organizaciones y se estima una demanda mínima de 250.000 de estos profesionales a medio plazo en España, según MBIT School, único centro de formación español dedicado exclusivamente a Business Intelligence y Big Dat.

Según Burtch Works,
el 32% de los científicos de datos en activo vienen del mundo de las matemáticas y la estadística,
el 19% de la ingeniería informática
el 16% de otras ingenierías.


Emerging Technology Hype Cycle for 2016
Machine Learning
Cognitive Expert Advisors
Software-Defined Security
Connected Home
Autonomous Vehicles
Blockchain
Nanotube Electronics
Smart Robots
Software-Defined Anything (SDx)
Micro Data Centers
Gesture Control Devices
IoT Platform
Commercial UAVs (Drones)
Affective Computing
Smart Data Discovery
Virtual Personal Assistants
Brain-Computer Interface
Natural-Language Question Answering
Conversational User Interfaces
Volumetric Displays
Enterprise Taxonomy and Ontology Management
Smart Workspace
Personal Analytics
Human Augmentation
Quantum Computing
Virtual Reality
Data Broker PaaS (dbrPaas)
Augmented Reality
Neuromorphic Hardware
Context Brokering
802.11ax
General-Purpose Machine Intelligence
4D Printing
Smart Dust
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
Expectations
As of July 2016
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
Source: Gartner
© 2016 Gartner, Inc. and/or its affiliates. All rights reserved.
Gartner

	Mínimo	Máximo
España	250,000	500,000
Perú	36,567	73,134

01. ¿QUÉ ES UN CIENTÍFICO DE DATOS?

Contexto

TIBCO®

INGENIERO

Solucionar el problema del negocio.

IT / DESARROLLADOR

Maneja la Infraestructura.

ESTADÍSTICO

Construye el Modelo.

Ciencia de Datos

01. ¿QUÉ ES UN CIENTÍFICO DE DATOS?

Contexto

INGENIERO

Solucionar el problema del negocio.

IT / DESARROLLADOR

Maneja la Infraestructura.

ESTADÍSTICO

Valida el Modelo.

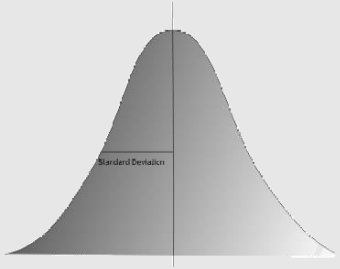
Ciencia de Datos

02. ANALICE Y ACTÚE EN ´MOMENTOS CRÍTICOS DEL NEGOCIO´

En tiempo real

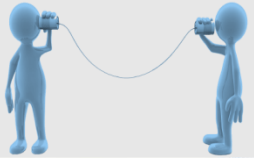
Ciencia de Datos

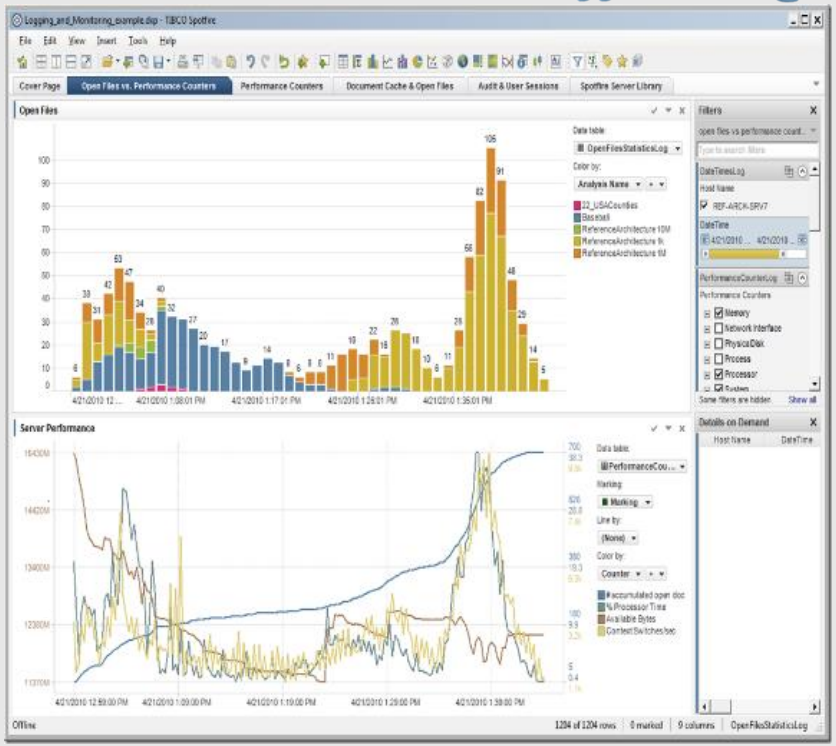
02. ANALICE Y ACTÚE EN ´MOMENTOS CRÍTICOS DEL NEGOCIO´
En tiempo real

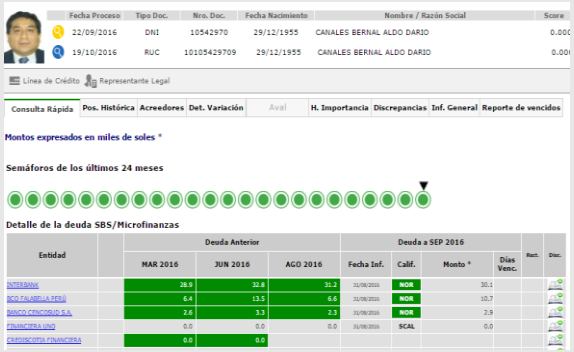



```
def add5(x):
   return x+5

def dotwrite(ast):
   nodename = getNodename()
   label=symbol.sym_name.get(int(ast[0]),ast[0])
   print '    %s [label="%s' % (nodename, label),
   if isinstance(ast[1], str):
      if ast[1].strip():
         print '= %s"];' % ast[1]
      else:
         print '"]'
   else:
      print '"];'
      children = []
      for n, child in enumerate(ast[1:]):
         children.append(dotwrite(child))
      print '    %s -> {' % nodename,
      for name in children:
         print '%s' % name,
```







Fecha Proceso	Tipo Doc.	Nro. Doc.	Fecha Nacimiento	Nombre / Razón Social	Score
22/09/2016	DNI	10542970	29/12/1955	CANALES BERNAL ALDO DARIO	0.000
19/10/2016	RUC	10105429709	29/12/1955	CANALES BERNAL ALDO DARIO	0.00

Línea de Crédito | Representante Legal

Consulta Rápida | Pos. Histórica | Acreedores | Det. Variación | Aval | H. Importancia | Discrepancias | Inf. General | Reporte de vencidos

Montos expresados en miles de soles *

Semáforos de los últimos 24 meses

Detalle de la deuda SBS/Microfinanzas

Entidad		Deuda Anterior			Deuda a SEP 2016				Rect.	Disc.
		MAR 2016	JUN 2016	AGO 2016	Fecha Inf.	Calif.	Monto *	Días Venc.		
INTERBANK		28.9	32.8	31.2	31/08/2016	NOR	30.1			
BCO FALABELLA PERU		6.4	13.5	6.6	31/08/2016	NOR	10.7			
BANCO CENCOSUD S.A.		2.6	3.3	2.3	31/08/2016	NOR	2.9			
FINANCIERA UNO		0.0	0.0	0.0	31/08/2016	SCAL	0.0			
CREDISCOTIA FINANCIERA		0.0	0.0							

MARTES 27 DE SEPTIEMBRE DEL 2016 | 21:38

Presentan App que mide riesgo crediticio personal

La solución tecnológica ha sido desarrollada por Fit pero es lanzada en sociedad con la empresa Deloitte

ESTÁS EN > PORTADA > TU DINERO >

Conozca una alternativa para reducir la tasa de interés de su crédito

Viernes, 30 de septiembre del 2016

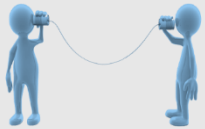
TU DINERO

05:01

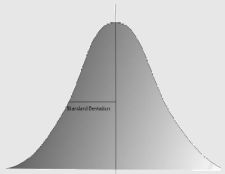
Especial TU DINERO. La tecnología está cada vez más arraigada en la vida de las personas. Sepa en la siguiente nota cómo el uso de un aplicativo puede ayudar a mejorar las condiciones de su préstamo.

```
def add5(x):
   return x+5

def dotwrite(ast):
   nodename = getNodename()
   label=symbol.sym_name.get(int(ast[0]),ast[0])
   print '    %s [label="%s' % (nodename, label),
   if isinstance(ast[1], str):
      if ast[1].strip():
         print '= %s"];' % ast[1]
      else:
         print '"]'
   else:
      print '"];'
      children = []
      for n, child in enumerate(ast[1:]):
         children.append(dotwrite(child))
      print '    %s -> {' % nodename,
      for name in children:
         print '%s' % name,
```

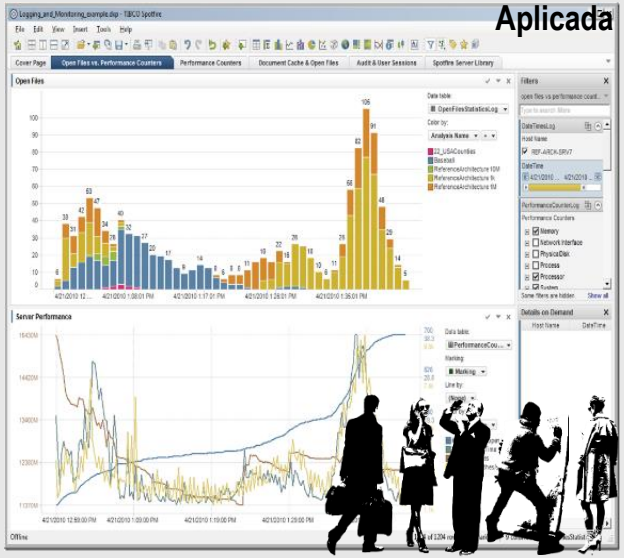


Comunicación



Programación

Estadística Aplicada

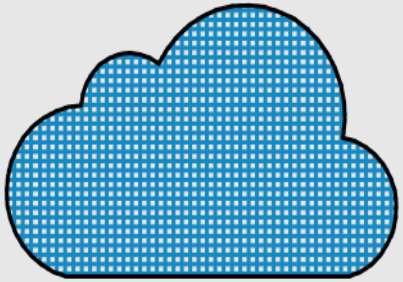


Visualización de Datos

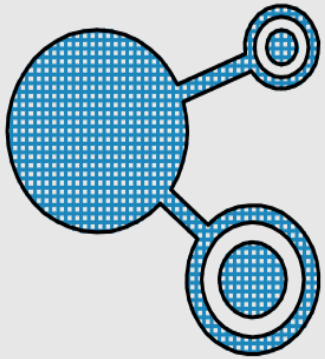
Experiencia en el Campo

03. PROCESO EN EL ANÁLISIS DE DATOS

Cinco Pasos

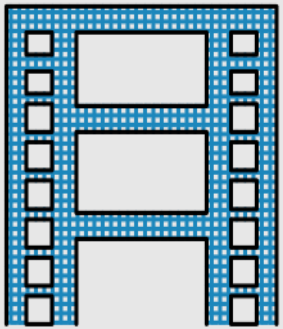


1/ Adquisición de Datos

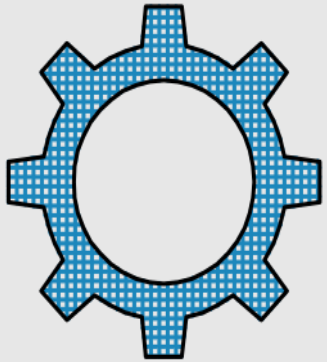


2/ Mineria de Datos (Data Munging)

3/ Exploración de Datos



4/ Generar Modelos



5/ Validación de Modelos

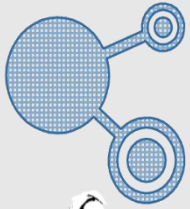
03. PROCESO EN EL ANÁLISIS DE DATOS

Cinco Pasos

TIBCO®

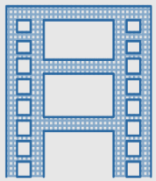
1 / Adquisición de Datos

2 / Minería de Datos (Data Wrangling)

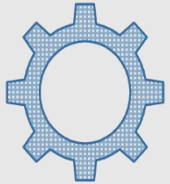


3 / Exploración de Datos

4 / Generar Modelos



5 / Validación de Modelos



Ciencia de Datos

04. ADQUISICIÓN DE DATOS

Paso Uno

UNO / Adquisición de Datos

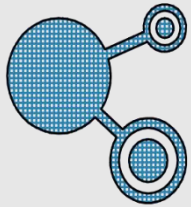
Databricks Certified Application

C i e n c i a d e D a t o s

05. MINERÍA DE DATOS

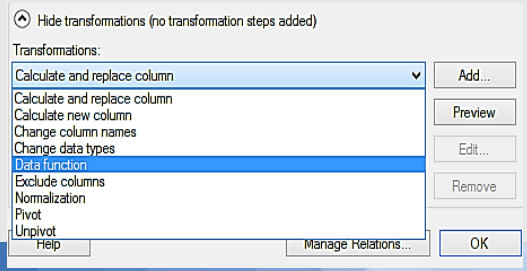
Paso Dos

DOS / Minería de Datos (Data Wrangling)



Data Wrangling o Data Mining

Minería de Datos: es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.



C i e n c i a d e D a t o s

ESTRATEGIA DE PRECIOS

Categoría	Productos con Stock	Productos con Matches						Productos con Matches con Stock			Ratio
		Totales		Sin Stock		Con Stock		Estoy más barato	Precio igualado	Estoy más caro	
Tv y Video	386	291	75%	71	18%	220	57%	46 / 21%	81 / 37%	93 / 42%	1,03
Electrohogar	978	786	80%	237	24%	549	56%	257 / 47%	118 / 21%	174 / 32%	0,97
Cómputo	349	244	70%	107	31%	137	39%	40 / 29%	32 / 23%	65 / 47%	1,10
Más	263	42	16%	31	12%	11	4%	3 / 27%	4 / 36%	4 / 36%	1,15
Entretenimiento	770	514	67%	172	22%	342	44%	118 / 35%	103 / 30%	121 / 35%	1,01
Muebles	625	79	13%	30	5%	49	8%	19 / 39%	22 / 45%	8 / 16%	0,95
CALZADO Y RELOJES	1355	525	39%	143	11%	382	28%	38 /	56 / 15%	288 / 75%	1,29
Belleza y Accesorios	719	335	47%	77	11%	258	36%	49 / 19%	164 / 64%	45 / 17%	0,99
Ripley Home	3724	857	23%	199	5%	658	18%	162 / 25%	402 / 61%	94 / 14%	0,99
Celulares	381	244	64%	67	18%	177	46%	41 / 23%	38 / 21%	98 / 55%	1,16
Deporte	1452	499	34%	227	16%	272	19%	67 / 25%	133 / 49%	72 / 26%	1,06
Infantil	1938	946	49%	299	15%	647	33%	146 / 23%	197 / 30%	304 / 47%	1,13

05. EJEMPLO MINERÍA DE DATOS

Paso Dos

***DOS* / Minería de Datos (Data Wrangling)**

PARA MAÑANA!

???

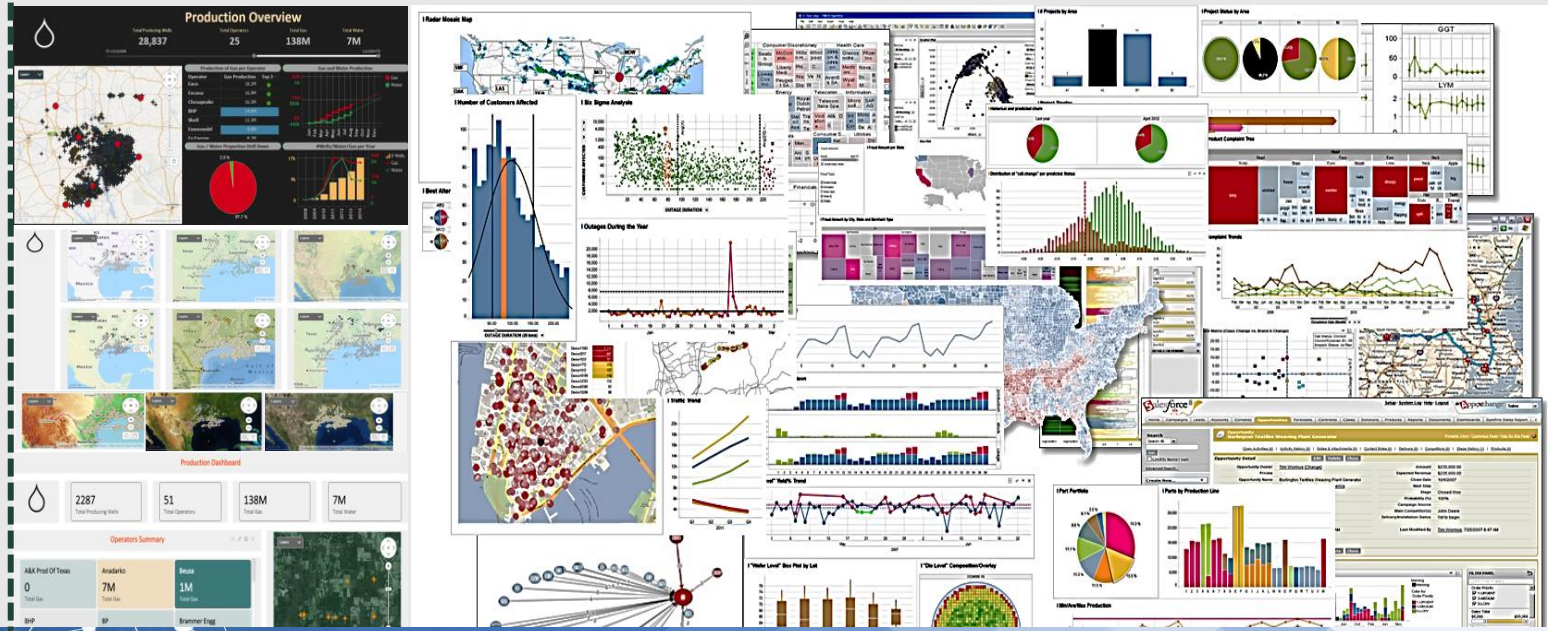
TIBCO®

CATEGORY	Q3-14	Q4-14	Q1-15	Q2-15	Q3-15	Q4-15	Q1-16	Q2-16	Well-Id
Total Cost in	2.00	2	3.00	4	2	2	3	6	Well 1
SMM	0.60	4	4.20	8	12	15	15	20	Well 1
Total Days	5.00	25	30.00	40	40	50	45	65	Well 1
Avg ABC per	6.00	50	40.00	60	100	130	120	150	Well 1
Operator	50.00	500	400.00	600	700	1000	500	600	Well 1
# of ABC	1.00	6	1.00	5	2	2	3	6	Well 2
Total ABC	0.60	4	4.20	8	12	15	15	20	Well 2
Total Cost in	5.00	25	30.00	40	40	50	45	65	Well 2

Ciencia de Datos

06. EXPLORACIÓN DE DATOS
Paso Tres

TRES / Exploración de Datos

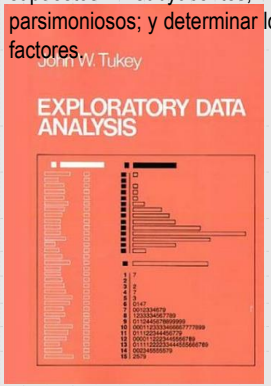


Ciencia de Datos

06. EXPLORACIÓN DE DATOS
Paso Tres

TRES /
Exploración de Datos

Análisis exploratorio de datos (EDA) es un enfoque / filosofía para el análisis de datos que emplea una variedad de técnicas (principalmente GRÁFICA) para maximizar la penetración de datos; descubrir la trayectoria subyacente; extraer variables importantes; detectar valores atípicos y anomalías; probar los supuestos subyacentes; desarrollar modelos parsimoniosos; y determinar los valores óptimos de los factores.



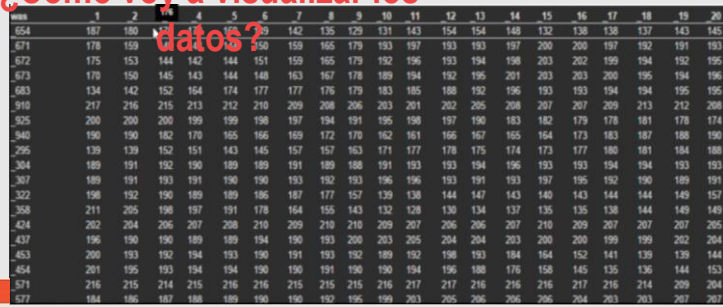
“El mayor valor de una imagen es cuando se nos obliga a darnos cuenta de lo que nunca esperábamos ver”.

06. EXPLORACIÓN DE DATOS
Paso Tres

TRES / Exploración de Datos

EXPLORACIÓN DE DATOS

¿Cómo voy a visualizar los datos?

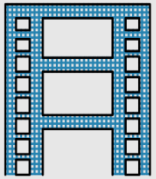


Ciencia de Datos

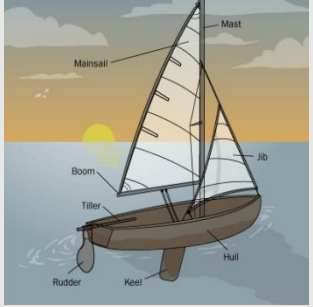
07. GENERAR MODELOS

Paso Cuatro

CUATRO / Generar Modelos



DATA RELATIONSHIPS



TIBCO®

- Numerical Vs Numerical
  - Lineal
  - Spearman R
- Numerical Vs Categorical
  - Anova
  - Kruskal-Wallis
- Categorical Vs Categorical
  - Chi-square

$$y = \beta_0 + \beta_1 x + \epsilon$$

Ciencia de Datos

El Científico de Datos

Aldo Canales Bernal
acanales@abab.com.pe