



# Introducción al análisis de datos

Semana - 1



# ¿Qué es?

- Data Science
- Data Mining
- Big Data
- Business Intelligence

## ¿Tienen algo en común?



# Definición de análisis avanzado de datos

Advanced Analytics is the autonomous or semi-autonomous examination of data or content using sophisticated techniques and tools, typically beyond those of traditional business intelligence (BI), to discover deeper insights, make predictions, or generate recommendations.

- Gartner Group

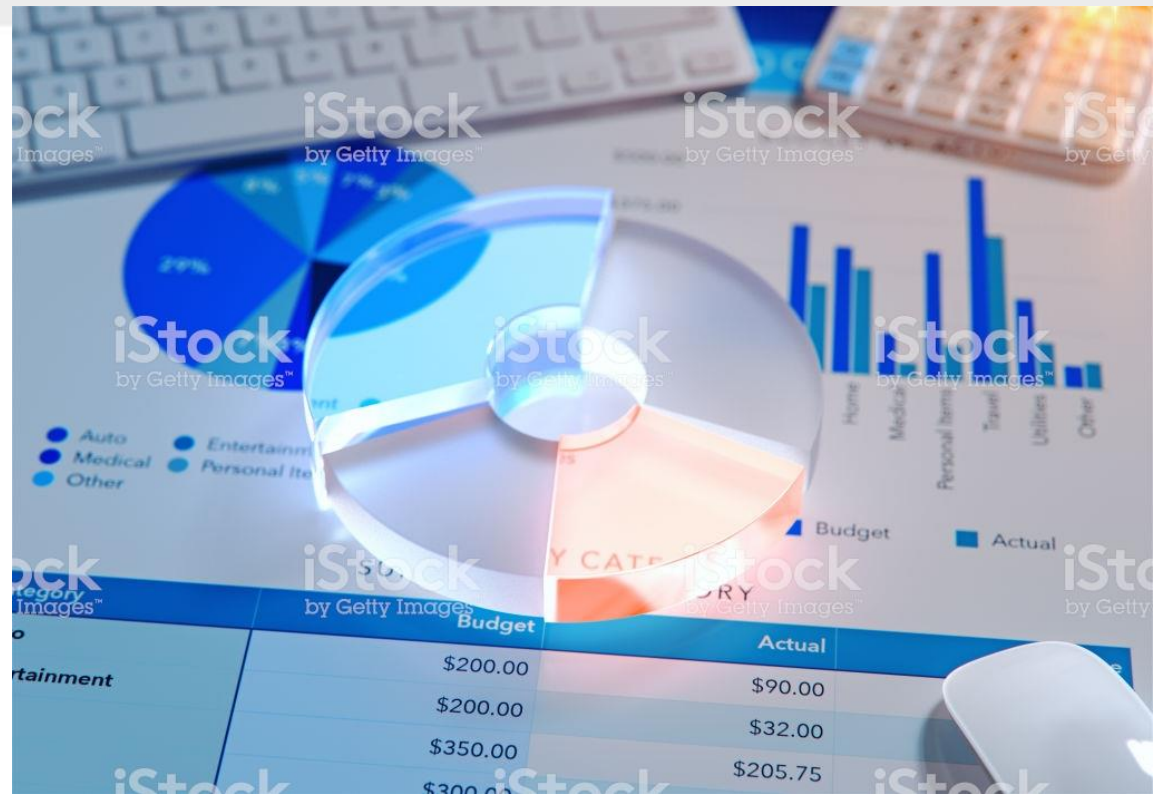
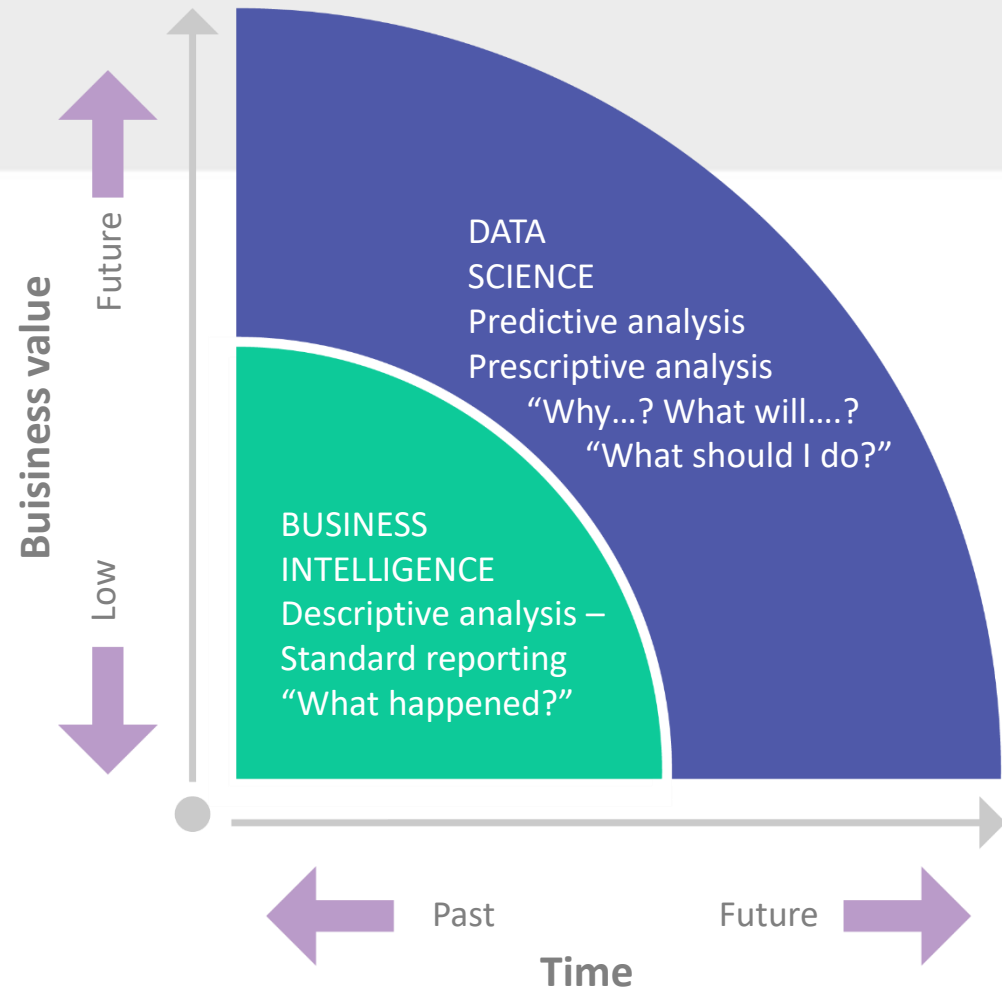


Foto tomada de: [https://www.istockphoto.com/photo/top-view-of-office-desk-with-data-chart-and-pie-chart-3d-rendering-gm925236050-253906983?irgwc=1&esource=AFF\\_IS\\_IR\\_SP\\_FreelImages\\_246195&asid=FreelImages&cid=IS&utm\\_medium=affiliate\\_SP&utm\\_source=FreelImages&utm\\_content=246195&clickid=SZazepUIPxYJR9RwUx0Mo3YyUKIQ3PX-dv3NSUO](https://www.istockphoto.com/photo/top-view-of-office-desk-with-data-chart-and-pie-chart-3d-rendering-gm925236050-253906983?irgwc=1&esource=AFF_IS_IR_SP_FreelImages_246195&asid=FreelImages&cid=IS&utm_medium=affiliate_SP&utm_source=FreelImages&utm_content=246195&clickid=SZazepUIPxYJR9RwUx0Mo3YyUKIQ3PX-dv3NSUO)



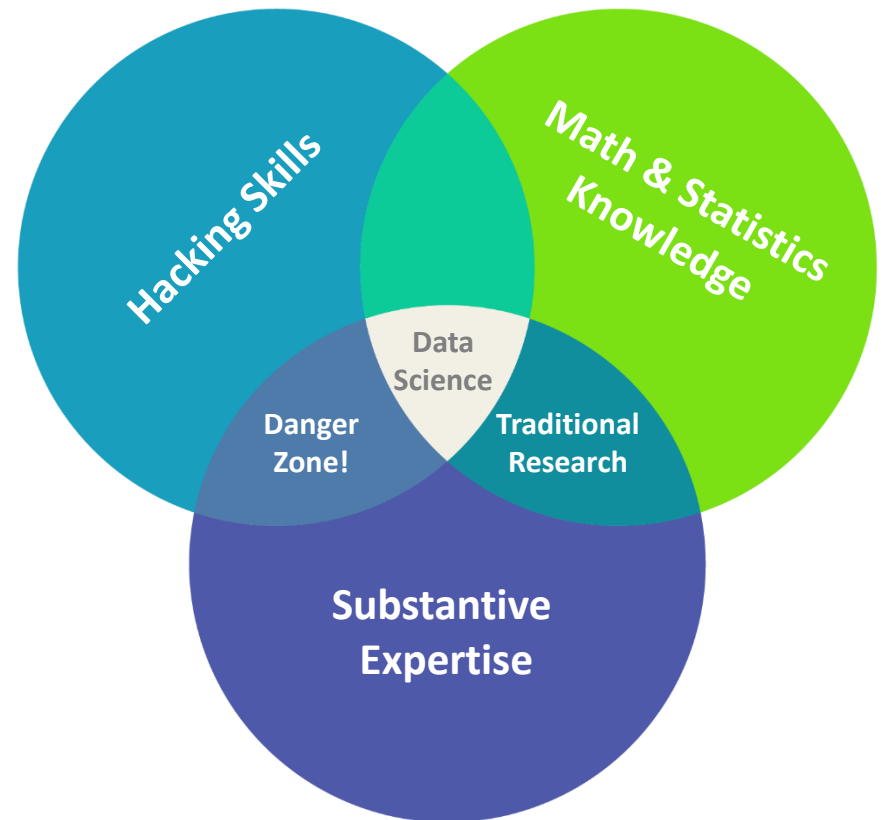
## BI vs Data Science

- La principal diferencia está en las preguntas que responde.
- BI responde a la pregunta de ¿qué pasó?
- Data Science responde a las preguntas
  - ¿Por qué?
  - ¿Qué pasará?
  - ¿Qué debo hacer?



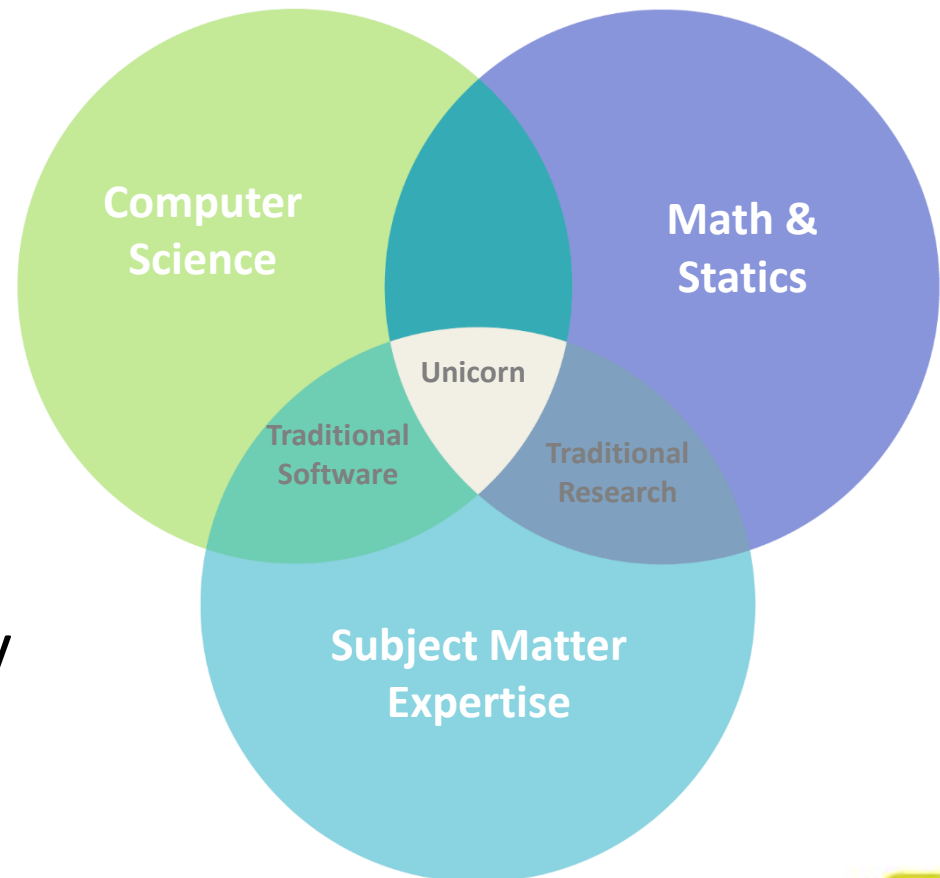
# Data Science, Diagrama de Venn

- Data science se compone de:
  - Hacking Skills – Habilidades de cómputo e informática
  - Math & Statistics – Habilidad con los números
  - Substantive Expertise – Dominio o expertise del tema
- ¿Por qué Danger Zone?



# Data Science, Diagrama real

- Seamos honestos... solo los unicornios tienen las tres cosas juntas.
- Por lo cual hay que concentrarse en los dos conjuntos superiores, el inferior llega solo a través de la experiencia y el tiempo (si llega...)



## DATA PREPARATION

### DATA CLEANING

INCONSISTENT DATATYPES  
MISSPELLED ATTRIBUTES  
MISSING AND DUPLICATE VALUES

### TRANSFORMATION



## EXPLORATORY DATA ANALYSIS



DEFINES AND REFINES  
THE SELECTION OF FEATURE  
VARIABLES THAT WILL BE USED  
IN THE MODEL DEVELOPMENT

## DATA MODELING

simplilearn

KNN



NAIVE BAYES

DECISION TREE

## VISUALIZATION AND COMMUNICATION

Tableau Power BI QlikView



# WHAT IS DATA SCIENCE?

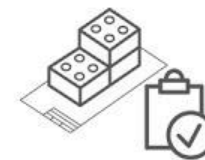
### DATA ACQUISITION

- WEB SERVERS
- LOGS
- DATABASES
- APIs
- ONLINE REPOSITORIES

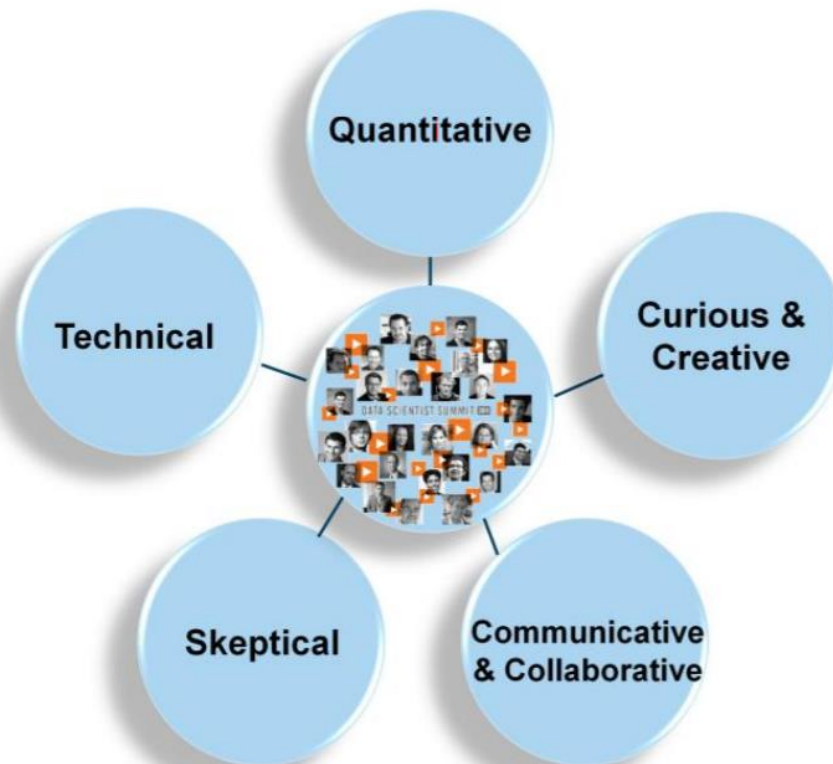
WHY?...WHY?...WHY?...



## DEPLOYS AND

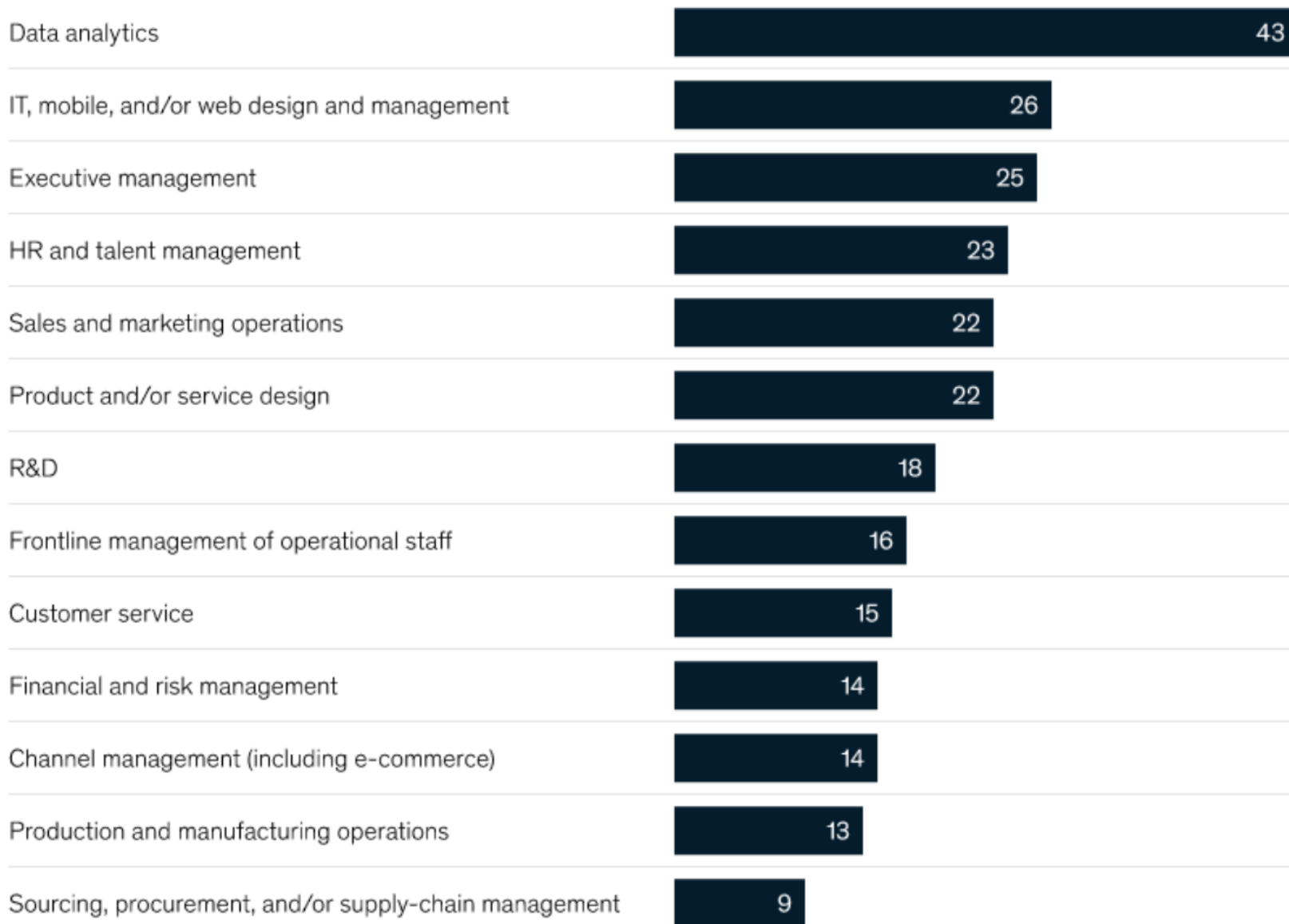


# ¿Quién es el Data Scientist?





## Business areas with greatest need to address potential skill gaps, % of respondents<sup>1</sup>



[A 2020 McKinsey survey](#) shows what business areas according to executives and managers will have the biggest potential skill gaps to address.

# Diferentes Roles



# Diferentes Roles

## DATA ANALYST DATA DETECTIVE

### Role

*Collects, processes and performs statistical data analyses*

### Mindset

*Intuitive data junkie with high "figure-it-out" quotient*



### Languages

*R, Python, HTML, Javascript, C/C++, SQL*

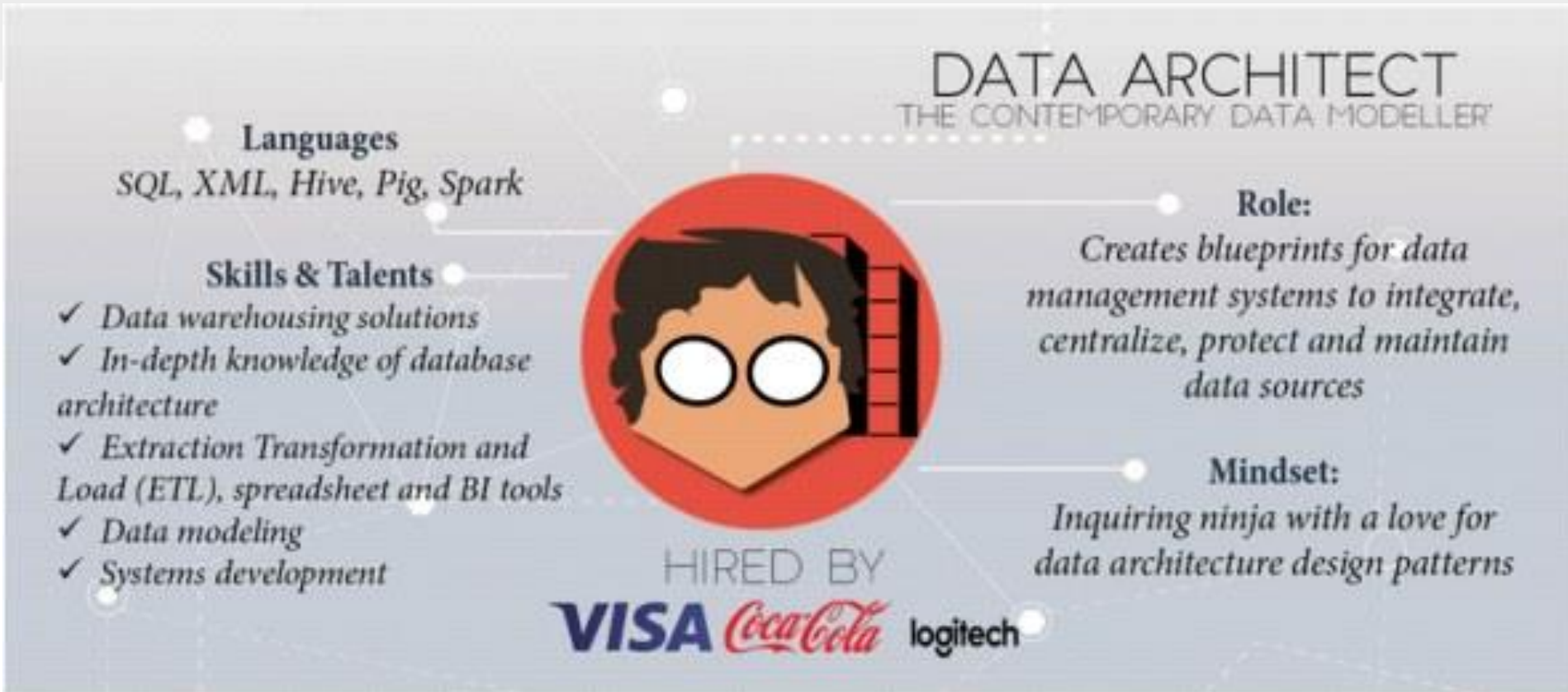
### Skills & Talents

- ✓ Spreadsheet tools (e.g. Excel)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Communication & visualization
- ✓ Math, Stats, Machine Learning

HIRED BY



# Diferentes Roles





# Diferentes Roles

## DATA ENGINEER

SOFTWARE ENGINEERS BY TRADE

### Role

*Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)*

### Mindset

*All-purpose everyman*



HIRED BY



### Languages

SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

### Skills & Talents

- ✓ Database systems (SQL & NO SQL based)
- ✓ Data modeling & ETL tools
- ✓ Data APIs
- ✓ Data warehousing solutions

# Diferentes Roles

## DATABASE ADMINISTRATOR 'DATABASE CARETAKER'

### Role

*Ensures that the database is available to all relevant users, is performing properly and is being kept safe*

### Mindset

*Master of Disaster Prevention*



HIRED BY



### Languages

*SQL, Java, Ruby on Rails, XML, C#, Python*

### Skills & Talents

- ✓ Backup & recovery
- ✓ Data modeling and design
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Data security
- ✓ ERP & business knowledge

# Diferentes Roles

## DATABASE ADMINISTRATOR 'DATABASE CARETAKER'

### Role

*Ensures that the database is available to all relevant users, is performing properly and is being kept safe*

### Mindset

*Master of Disaster Prevention*



HIRED BY



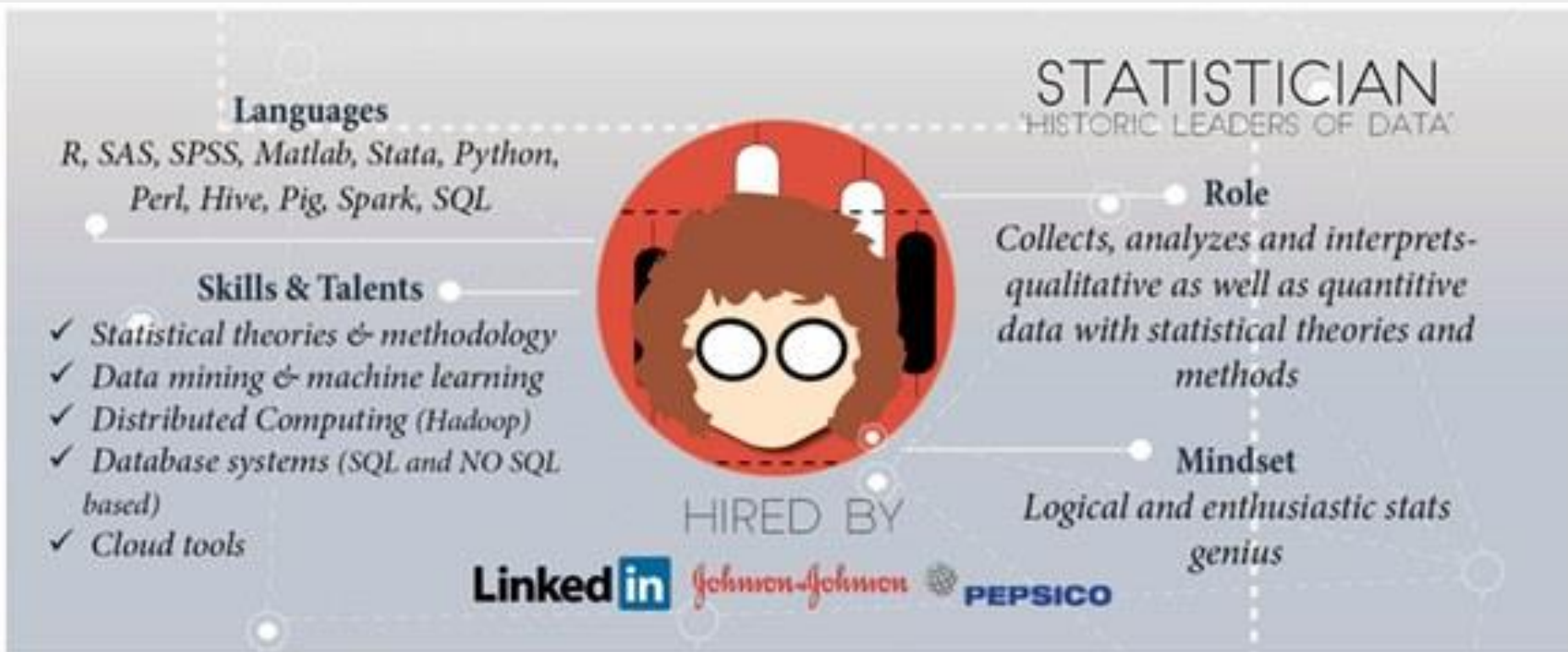
### Languages

*SQL, Java, Ruby on Rails, XML, C#, Python*

### Skills & Talents

- ✓ Backup & recovery
- ✓ Data modeling and design
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Data security
- ✓ ERP & business knowledge

# Diferentes Roles





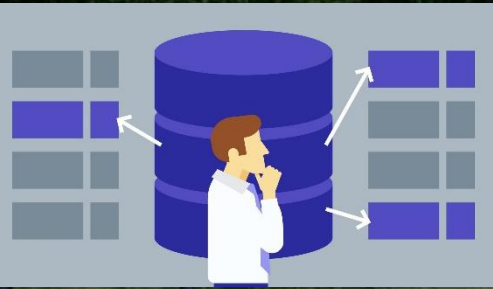
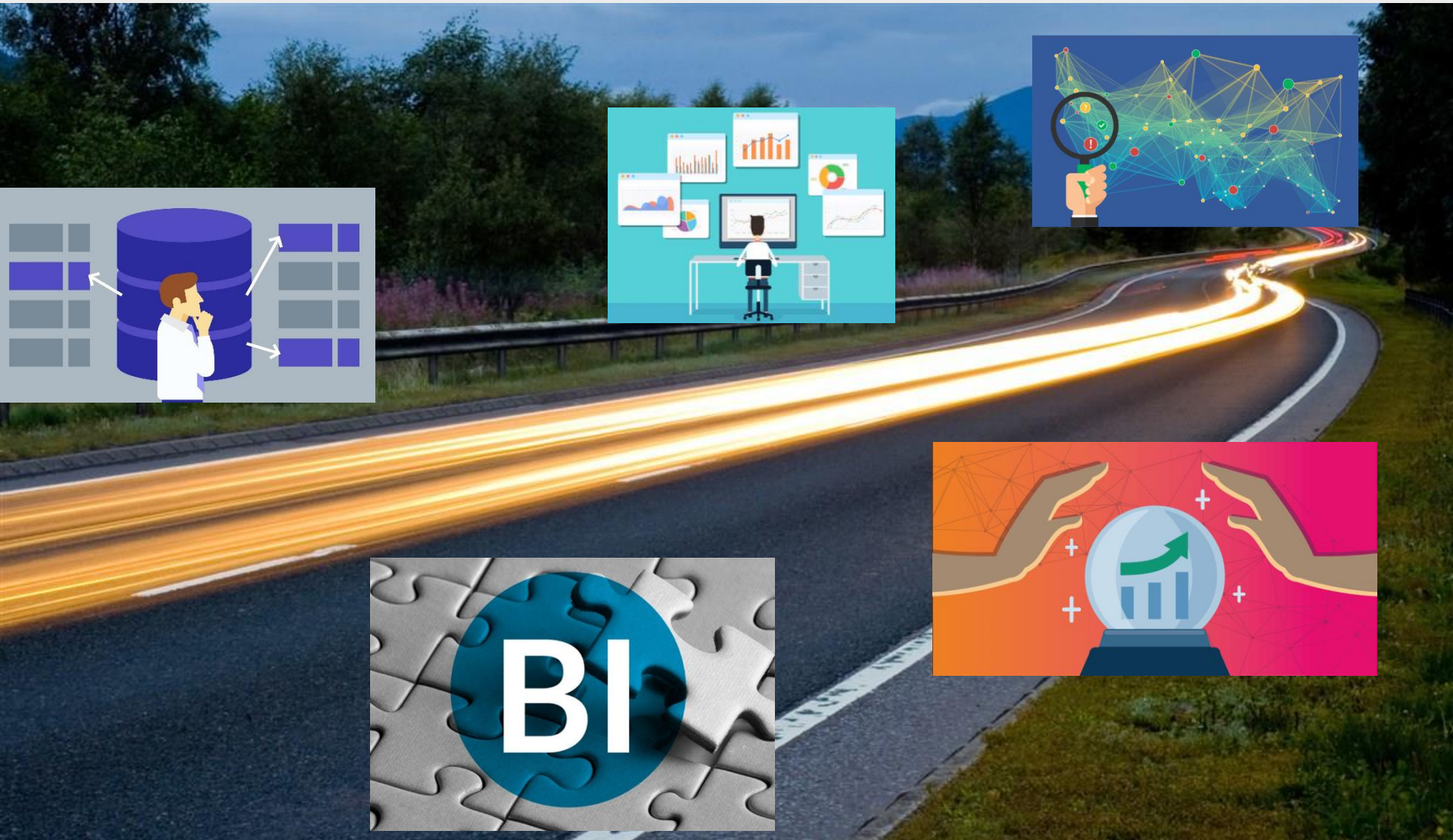
# Diferentes Roles



<https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html>



# Autopista de Data

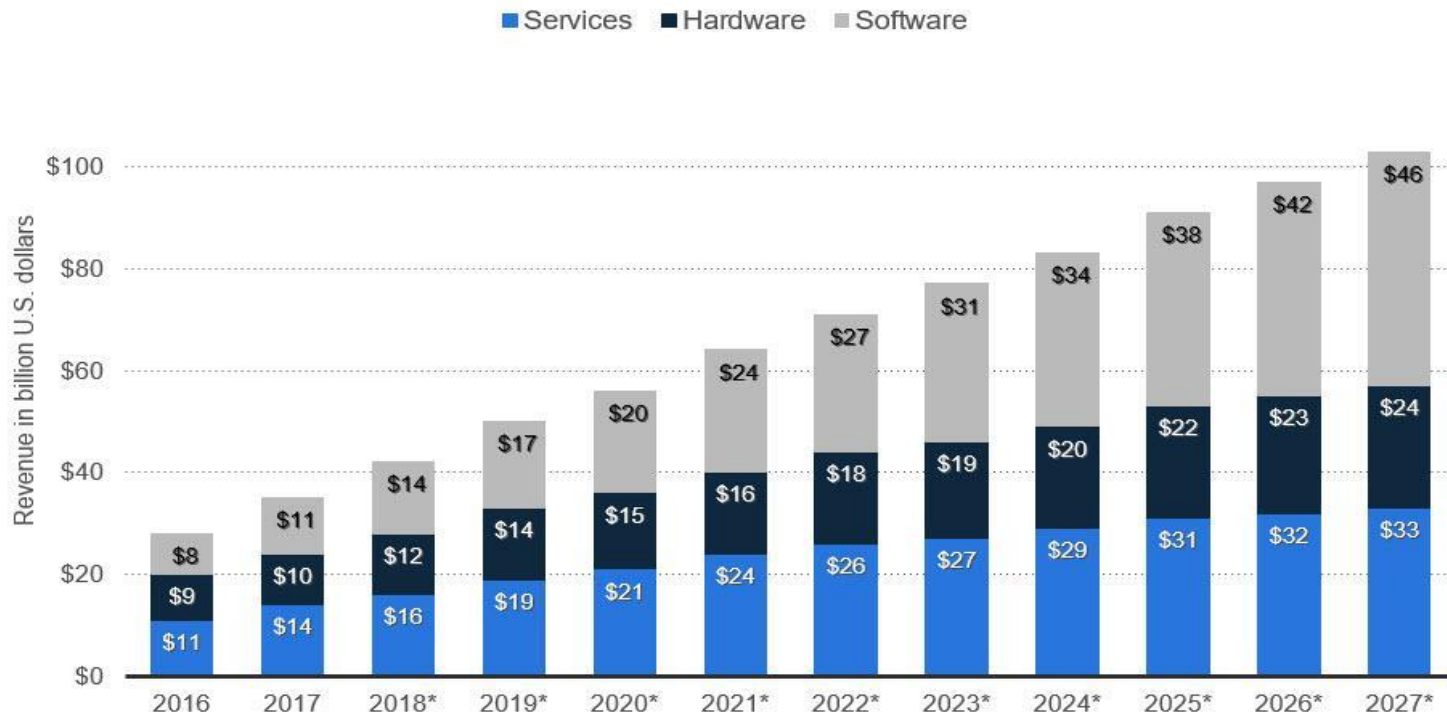






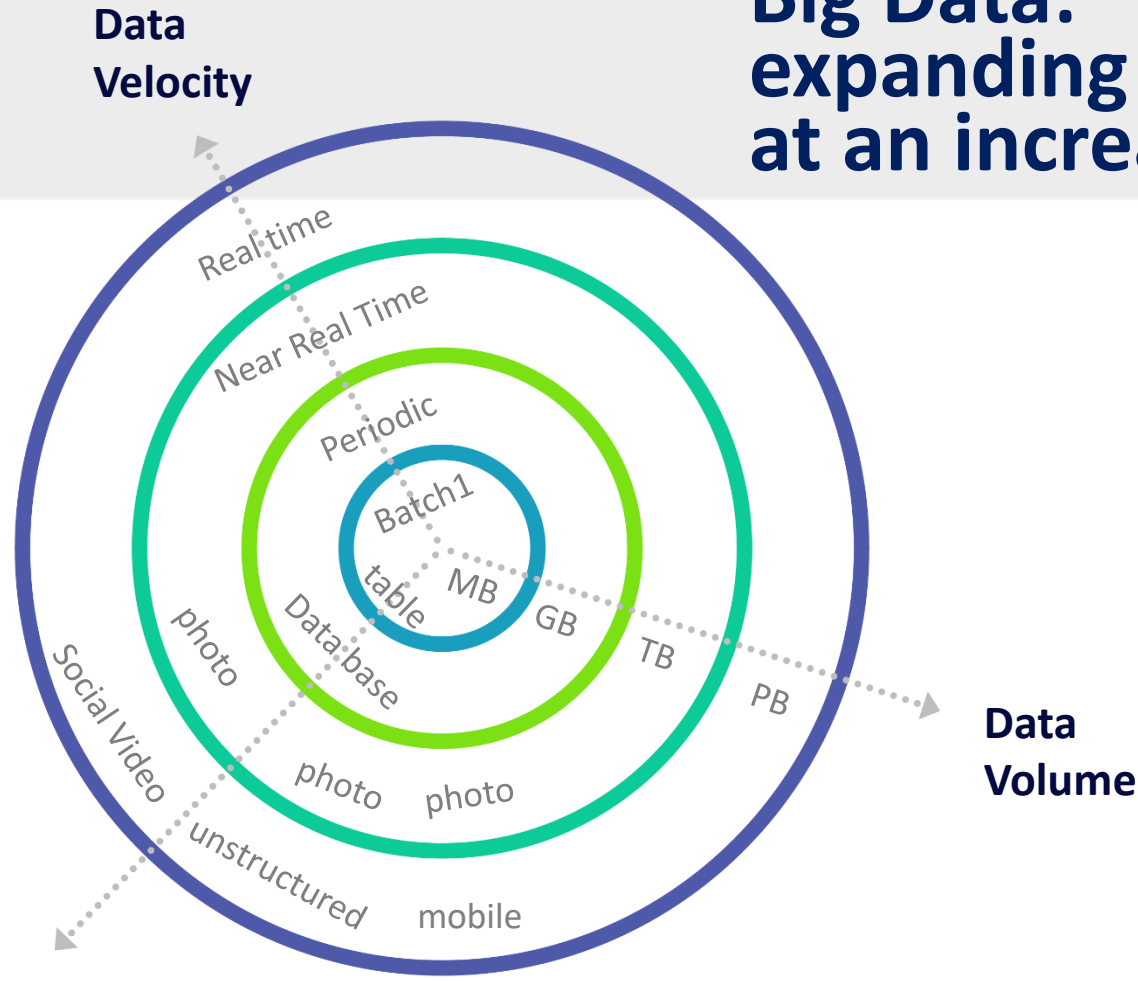
Global Big Data Revenue 2016-2027, by type

## Big Data Revenue Worldwide from 2016 to 2027, by major segment (in billion U.S. dollars)





# Big Data: expanding on 3 fonts at an increasing rate



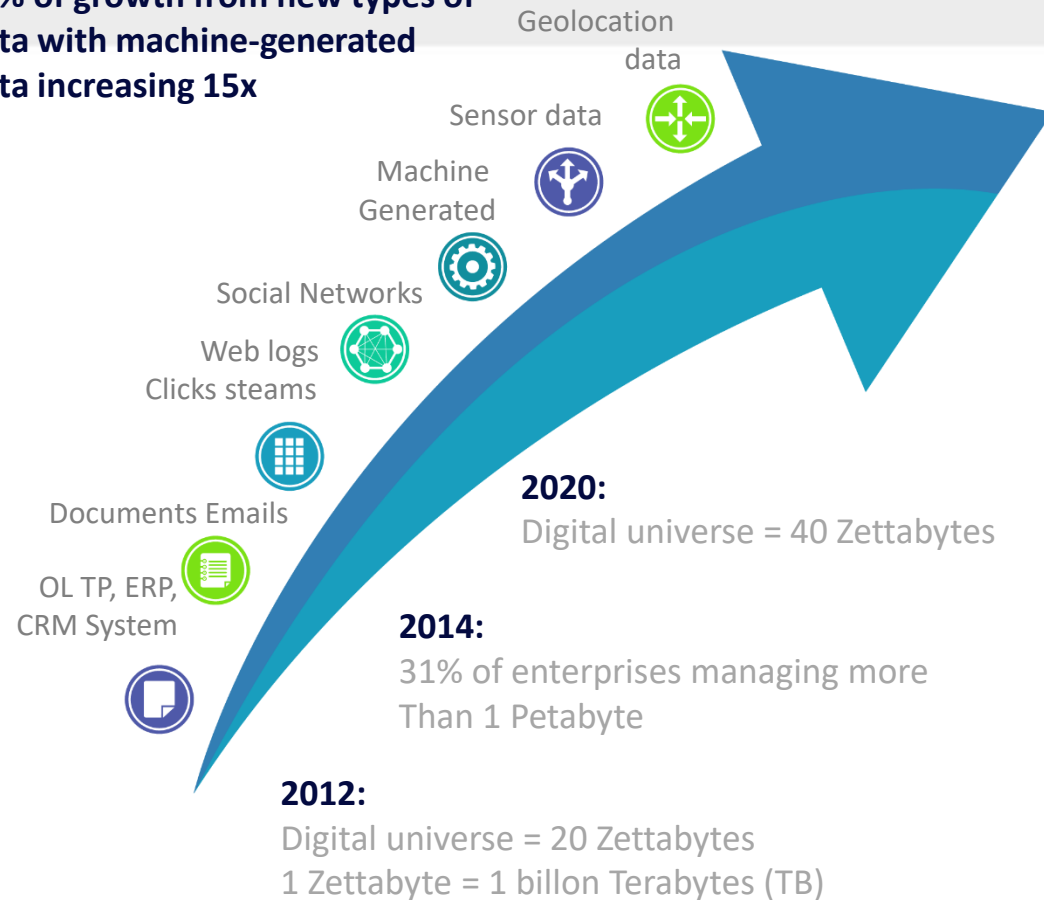
**Data  
Variety**

# Volumen de datos

- El volumen de datos continua creciendo exponencialmente.
- Últimamente una de las fuentes que produce dicho aumento proviene de las máquinas por medio del **IoT** y datos generados por máquinas y computadoras.

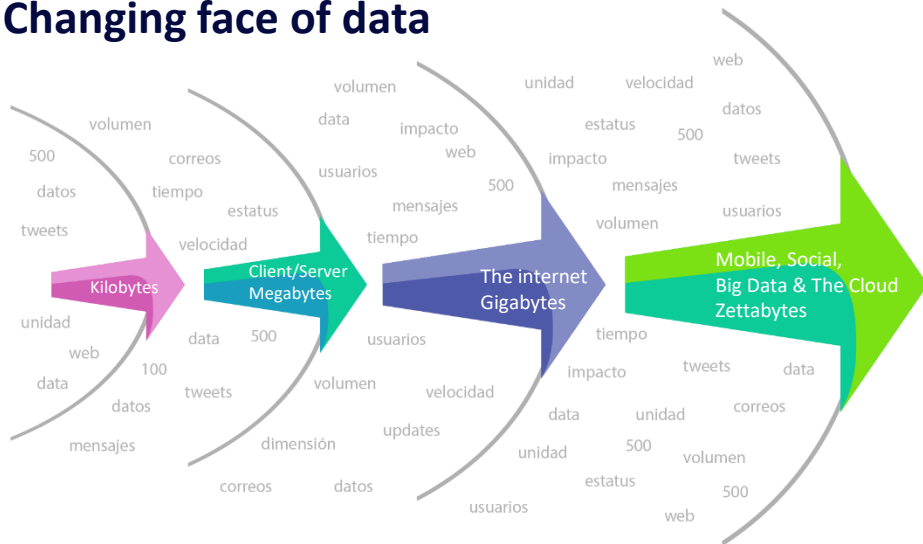
## Data Continues to Grow Sharply

85% of growth from new types of data with machine-generated data increasing 15x










# Volumen de los datos

## Changing face of data



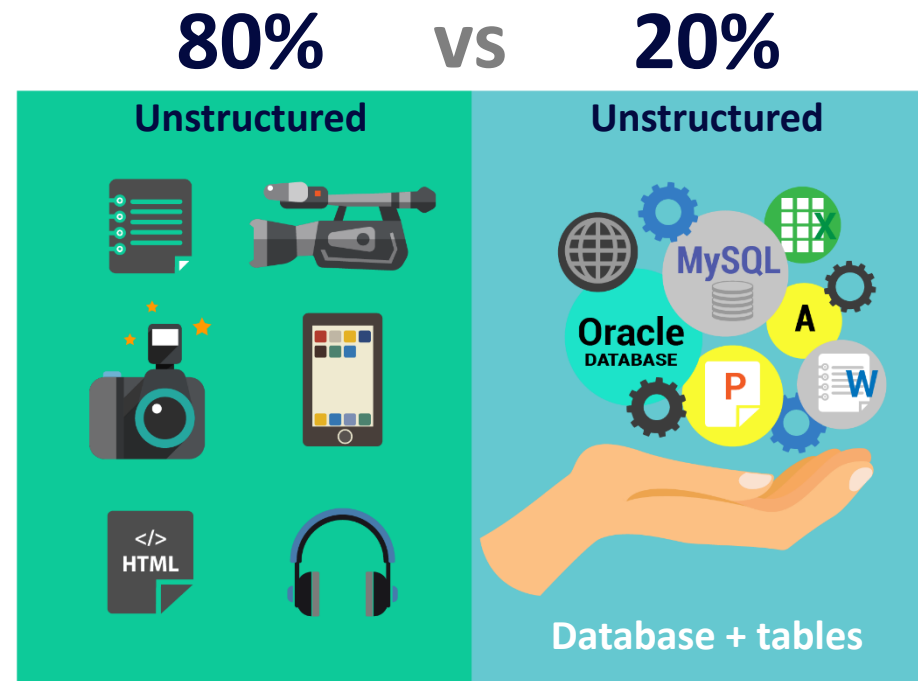
## Every 60 seconds

-  **98,000 + tweets**
-  **695,000 status updates**
-  **168 million+ email sent**
-  **11 million instant messages**
-  **698,445 google searches**
-  **1,820TB of data created**
-  **217 New mobile web users**

# Variedad de datos

- Anteriormente (antes del 2000) el volumen de datos más grande era generado por las grandes empresas y compañías con sistemas de información monolíticos.
- Hoy en día, más del 80% de los datos son no estructurados.

**Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003**



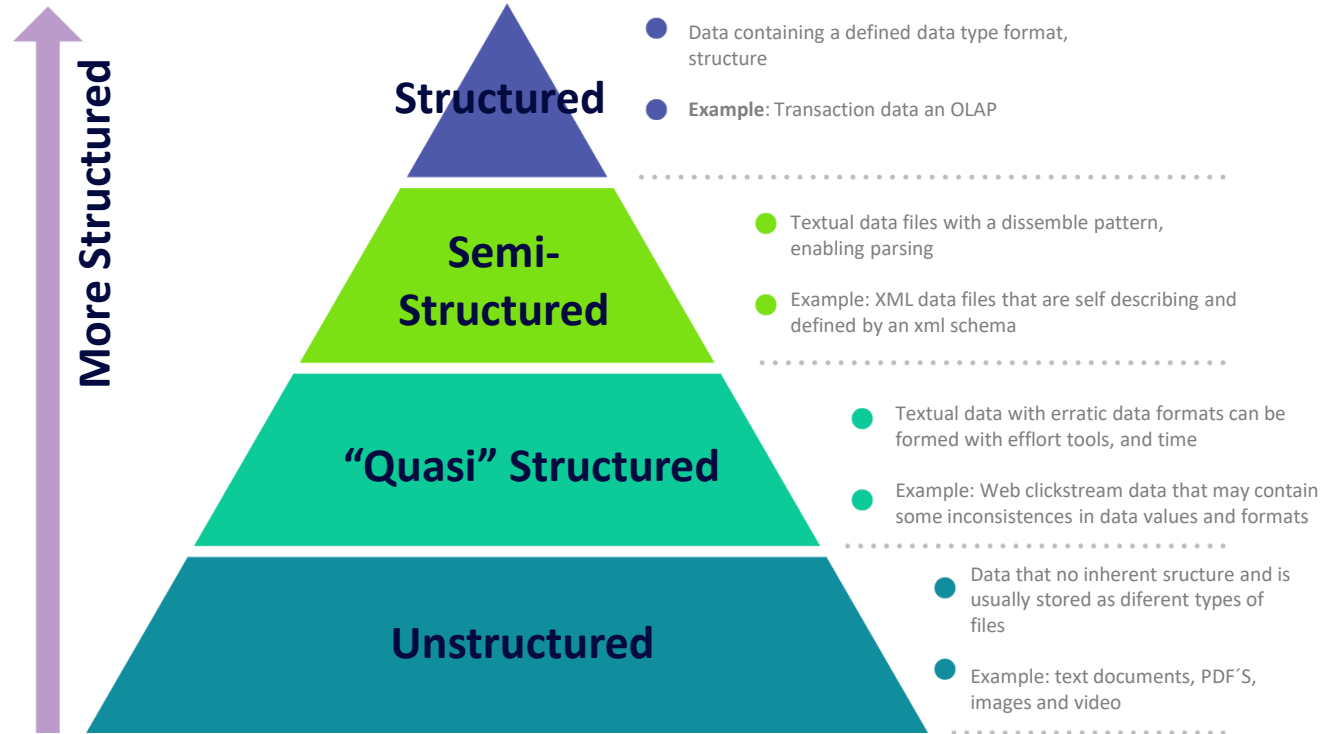


# Variedad de los datos

## BIG DATA CHARACTERISTICS: DATA STRUCTURES

Data Growth is Increasingly Unstructured

- Hoy en día el gran volumen de datos que podemos encontrar se basa principalmente en contenido de dato no estructurado.



Product code	Product name	Color	Price	Stock quantity
PCODE	PNAME	COLOR	PRICE	SQUANTITY
101L	Blouse	Blue	35.00	62
101M	Blouse	White	35.00	85
201M	Polo shirt	White	36.40	29
202M	Polo shirt	Red	36.40	67
302S	Skirt	White	51.10	65
353L	Skirt	Red	47.60	18
353M	Skirt	Green	47.60	56
411M	Sweater	Blue	84.00	12
412M	Sweater	Red	84.00	22
591L	Socks	Red	2.50	300
591M	Socks	Blue	2.50	90
591S	Socks	White	2.50	280
671L	Sweatsuit	White	45.00	45
671M	Sweatsuit	Blue	45.00	76

timestamp  
ndbox / Omniture.0.tsv.gz

IP Address









```

31799426      2012-03-15 01:17:06      28600057559854677
0      99.122.210.248      1      0
{7AAB8415-E803-3C5D-7100-E362D7F67CA7}
U      en-us,en;
Y      2      0      304      sbcglobal.net 1
1,10020,00007 Mozilla/5.0 (Windows; U; Windows NT 6.1;
0      2      3      0      homestead
0
  
```

Geocoded IP Address

## Tipos de datos

### data types

 Text files and documents	 Server, website and application logs	 Sensor data	 Images
 Video files	 Audio files	 Emails	 Social media data

```

<Customers>
- <Customer CustomerName="Arshad Ali" CustomerID="C001">
- <Orders>
- <Order OrderDate="2012-07-04T00:00:00" OrderID="10248">
<OrderDetail Quantity="5" ProductID="10"/>
<OrderDetail Quantity="12" ProductID="11"/>
<OrderDetail Quantity="10" ProductID="42"/>
</Order>
</Orders>
<Address> Address line 1, 2, 3</Address>
</Customer>
- <Customer CustomerName="Paul Henriot" CustomerID="C002">
- <Orders>
- <Order OrderDate="2011-07-04T00:00:00" OrderID="10245">
<OrderDetail Quantity="12" ProductID="11"/>
<OrderDetail Quantity="10" ProductID="42"/>
</Order>
</Orders>
<Address> Address line 5, 6, 7</Address>
</Customer>
- <Customer CustomerName="Carlos Gonzlez" CustomerID="C003">
- <Orders>
- <Order OrderDate="2012-08-16T00:00:00" OrderID="10283">
<OrderDetail Quantity="3" ProductID="72"/>
</Order>
</Orders>
<Address> Address line 1, 4, 5</Address>
  
```

# Velocidad de los datos

## Comparing High-Velocity Data & Big Data

### High-Velocity Data

- Real-time
- Performance & Volume Challenges
- Use Cases: Operations & Analytics

### Big Data

- Batch Process
- Volume Challenge
- Use Case: Analytics



- La velocidad sin embargo, es qué tan oportunos son los datos para analizar los mismos cuando estos se necesitan.
- Aunque el escenario idóneo es “tiempo-real”, esto implica contar con herramientas e infraestructura para garantizar un buen desempeño.



# THE 4 V'S OF BIG DATA

**40 ZETTABYTES**  
of data will be created by  
2020, an increase of 300  
times from 2005



**6 BILLION PEOPLE**  
have cell phones  
WORLD POPULATION: 7 BILLION



## Volume

SCALE OF DATA

**2.5 QUINTILLION BYTES**  
of data are created  
each day



Most companies in the  
U.S. have at least  
**100 TERABYTES**  
of data stored



As of 2011, the global size of  
data in healthcare was  
estimated to be  
**150 EXABYTES**



**30 BILLION  
PIECES OF CONTENT**  
are shared on facebook  
every month



## Variety

DIFFERENT  
FORMS OF DATA

**4 BILLION +  
HOURS OF VIDEO**  
are watched on  
YouTube each month



**4 MILLION TWEETS**  
are sent per day by about  
200 million monthly active  
users



## Velocity

ANALYSIS OF  
STREAMING DATA

The New York Stock  
Exchange captures  
**1TB OF TRADE  
INFORMATION**  
during each trading  
session



Modern cars have  
close to  
**100 SENSORS**  
that monitor items such as  
fuel level and tire pressure



**1 IN 3 BUSINESS  
LEADERS**  
don't trust the information  
they use to make  
decisions



## Veracity

UNCERTAINTY  
OF DATA

**27% OF RESPONDENTS**  
in one survey were unsure  
of how much of data  
was inaccurate



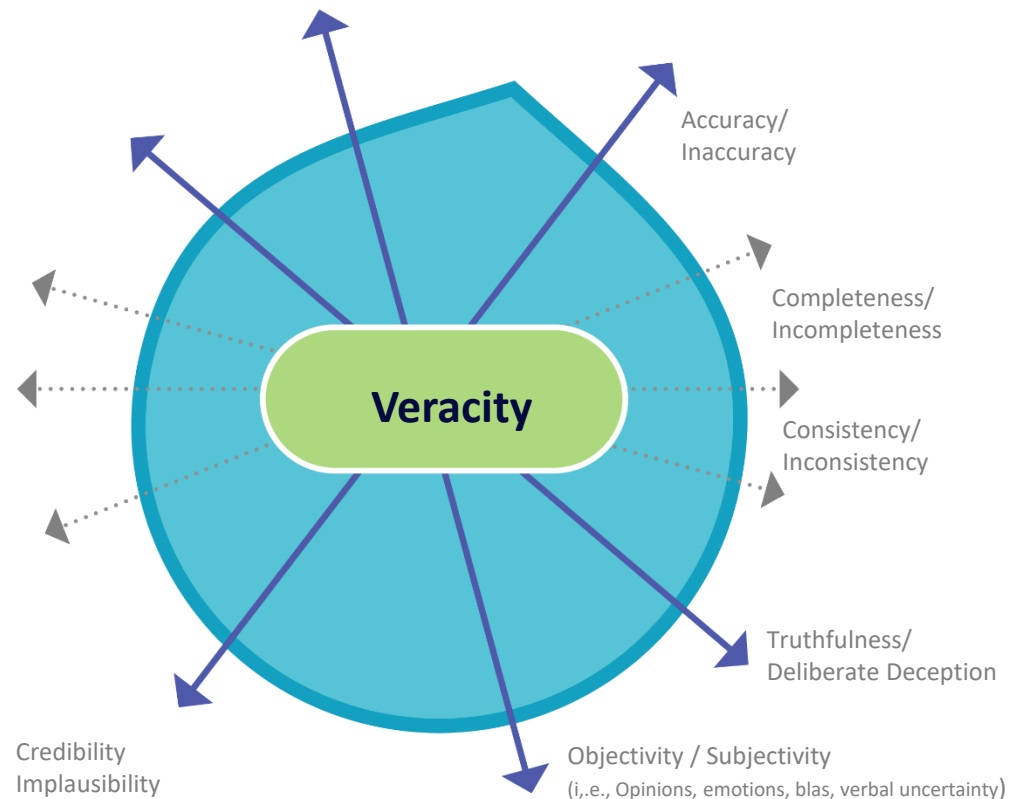


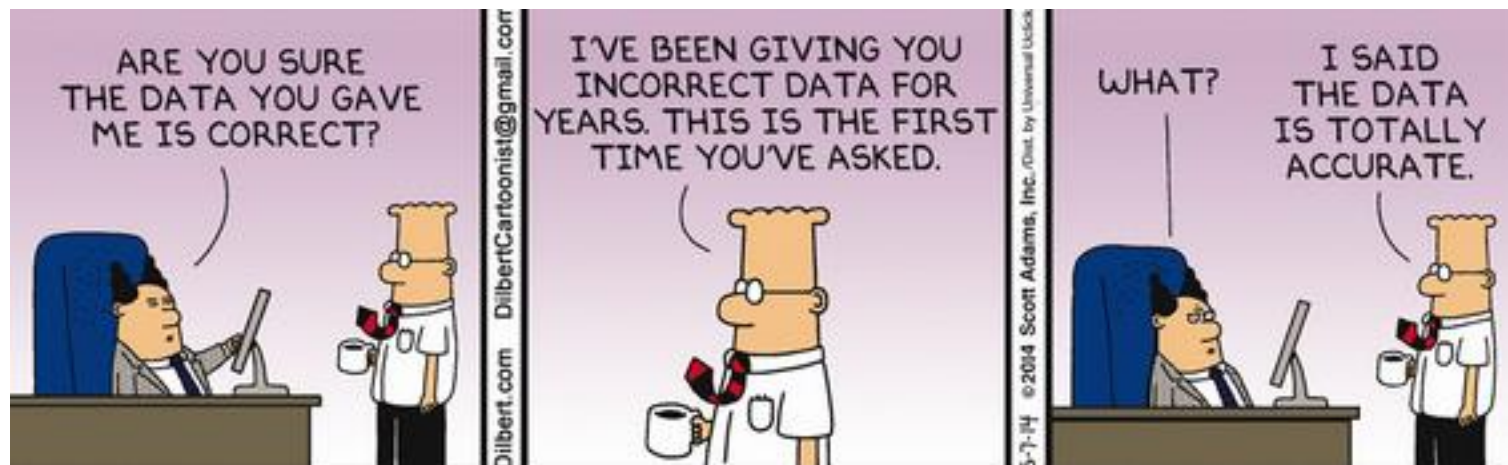
# Veracidad de los datos

- De nada sirve analizar datos y contar con herramientas para dicho análisis, si los datos carecen de veracidad.
- La falta de veracidad es un reto ya que involucra contar con procesos de validación y corrección de la información que permitan presentar únicamente la información de forma “limpia”.

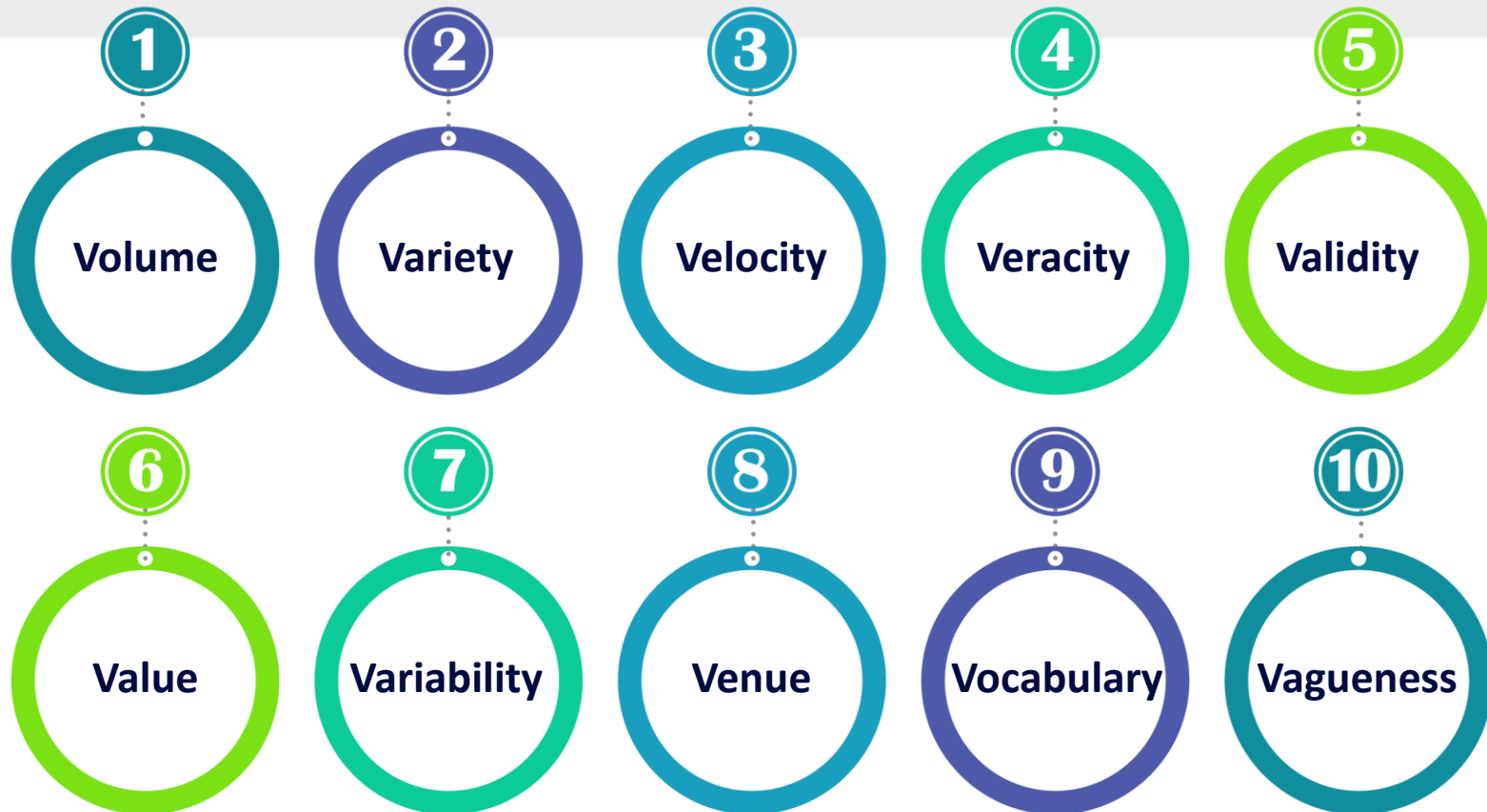
## Veracity Roadmap\*: Components of Big data 4<sup>th</sup> V

\*to be calculated as composite veracity index





## 10 Big Data vs



# Repositorios de Data





# Banco ABC

- El Banco ABC se encuentra en un proceso de cambio:
  - Evoluciona de un banco pequeño a un banco global.
  - Debe cambiar una infraestructura monolítica a una que soporte de mejor forma el análisis de datos.
  - Crece por medio de adquisiciones y compras de otros bancos.
  - Incrementa su cartera de clientes y servicios.
- Elabore una estrategia orientada al análisis de la información enfocándose en:
  - ¿Qué tipo de repositorio cree que tiene ABC? ¿Cuál necesitará?
  - ¿Cómo evolucionan sus necesidades de análisis de información?
  - ¿Qué necesitará desde el punto de vista del “Data Scientist”?

# ¿Qué busca una empresa al analizar los datos?



**Current Business Problems Provide Opportunities for Organizations to Become More Analytical & Data Drive**

Driver	Examples
<b>1</b> Desire to optimize business operations	Sales, pricing, profitability, efficiency
<b>2</b> Desire to identify business risk	Customer churn, fraud, default
<b>3</b> Predict new business opportunities	Upsell, cross-sell, best new customer prospect
<b>4</b> Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending Basel II

# Big Data en Healthcare



## Situación

- En ocasiones elegir el staff de un hospital puede ser complejo ya que si hay mucho staff esto eleva los costos y si hay poco esto puede repercutir en el servicio.



## Uso de Big Data

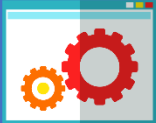
- En el Assistance Publique-Hôpitaux de Paris, data scientists crearon análisis de series de tiempo tomando información de hospitales alrededor del mundo que presentaran contextos similares (demografía, nivel socio económico, tendencias de emergencias, etc.)



## Resultado

- El resultado fue contar con modelos de análisis predictivo, que por medio de machine learning se adaptan constantemente, y a través de un navegador permiten a los médicos predecir el nivel de staff que necesitan cada hora.

# Big Data en Servicios Públicos



## Situación

- En 2017 la firma de analistas de tráfico Inrix dijo que Los Angeles era la peor ciudad de todo EEUU, por lo que la ciudad trabajó con Inrix para revertir esta situación.



## Uso de Big Data

- Junto con el Departamento de control de tráfico crearon una herramienta de captura de información por medio de GPS que identifica los puntos más críticos de la ciudad y los factores correspondientes a dichos puntos, tales como volumen vehicular, número de rutas alternas, presencia de puentes o pasos a desnivel, ancho de las carreteras, etc.

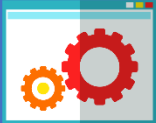


## Resultado

- Como resultado se obtiene ahora información en tiempo real que permite a los policías redireccionar el tráfico a puntos menos congestionados, priorizar en que rutas se debe invertir para ampliar carreteras o crear nuevas arterias vehiculares.



# Big Data en el clima



## Situación

- En los Estados Unidos, las pérdidas y daños causados por el clima y tempestades suma anualmente un promedio de \$500 millones.



## Uso de Big Data

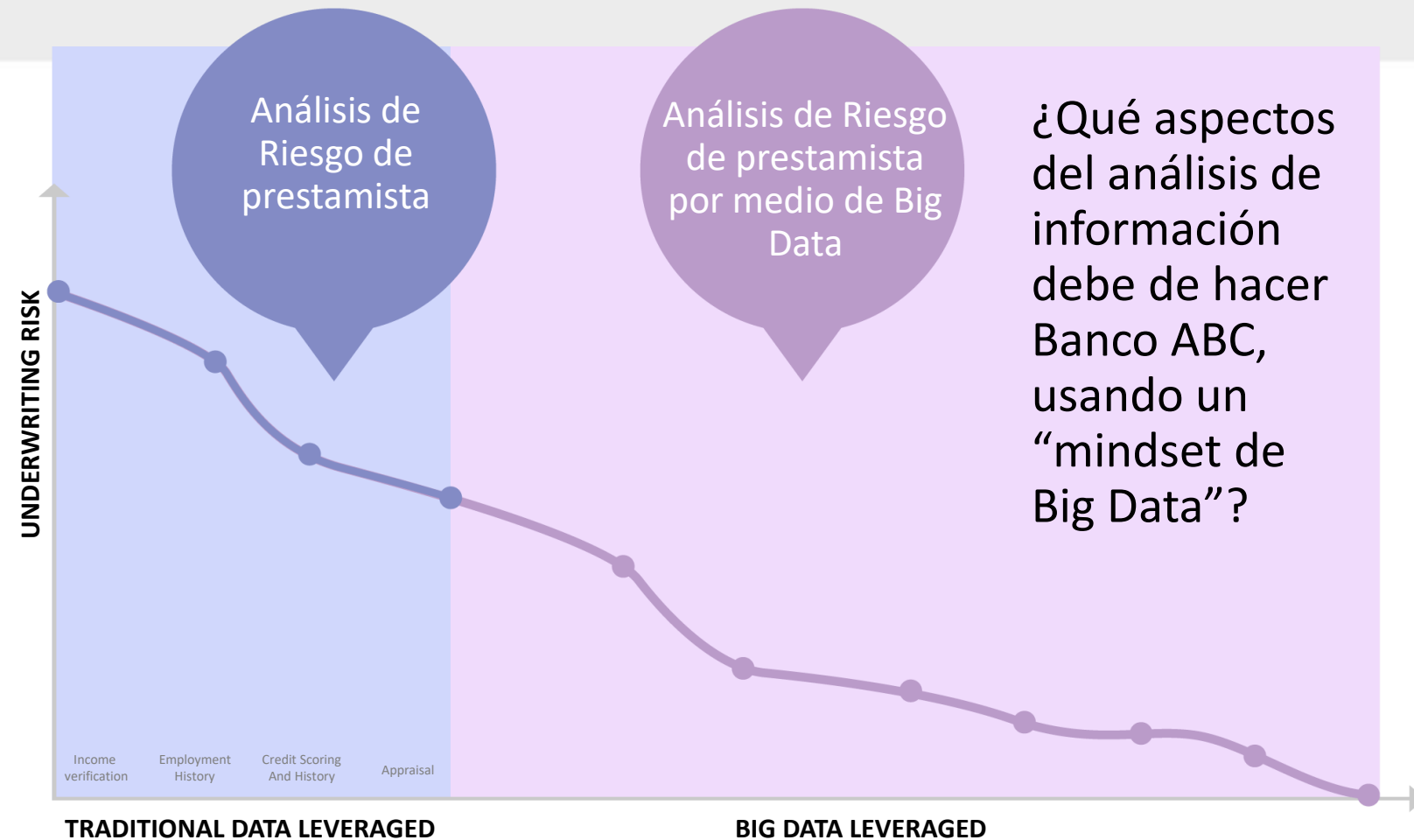
- IBM adquirió la compañía de “Wheater Forecasting and Information Technology” e introdujo más de 100,000 sensores para captura de información, así como también drones, apps y otros dispositivos.



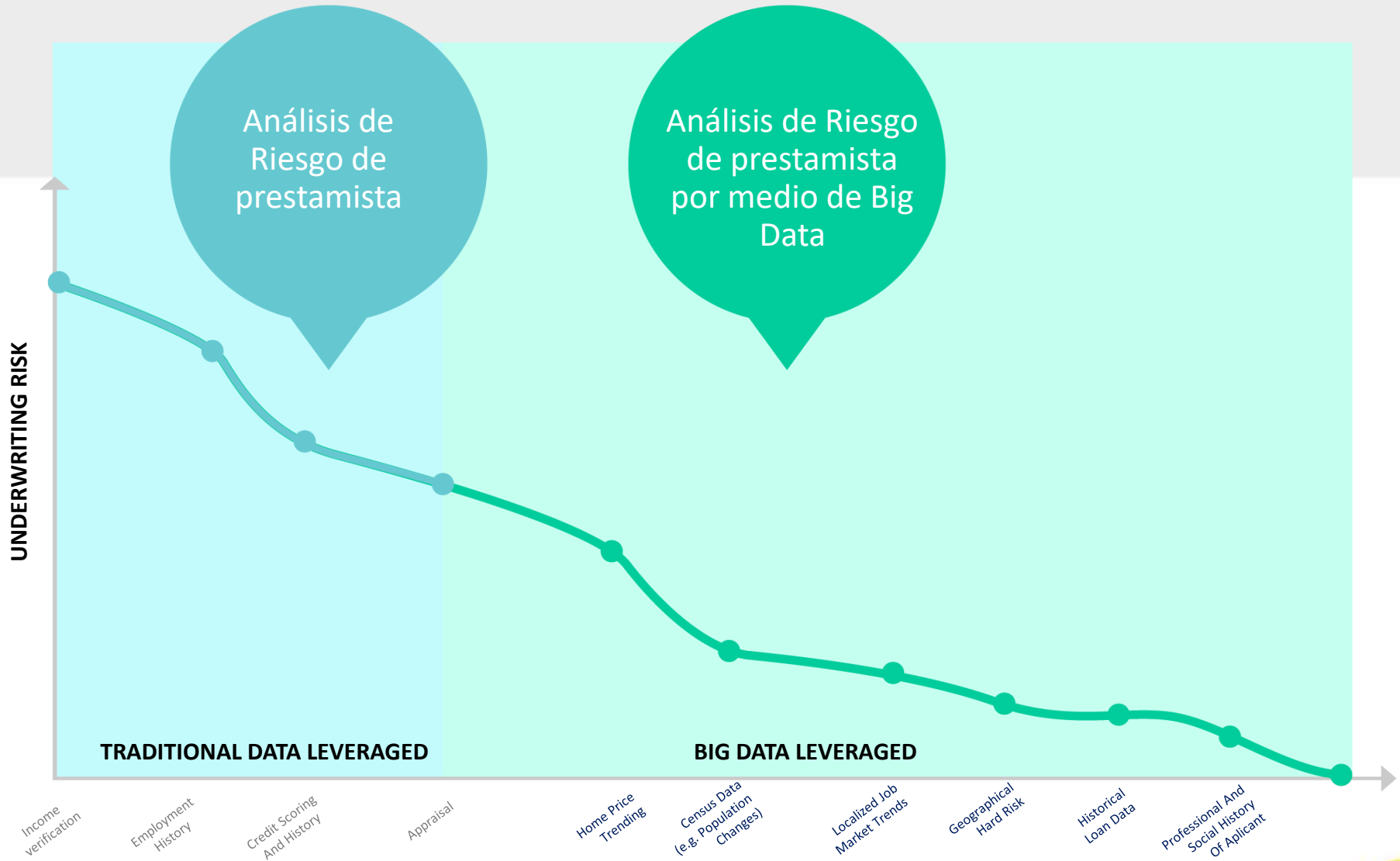
## Resultado

- Esto permite a IBM a través de Watson, su tecnología de Inteligencia Artificial, crear alertas en tiempo real para los negocios, empresas y el público en general que potencialmente pueden causar daños en sus áreas geográficas.

# Continuación Banco ABC



## Continuación Banco ABC



# Tarea: Análisis de aplicación Big Data

- Ingrese a <http://bigdata.stratebi.com/>
- Diríjase al link de “Demos”
- Realice un análisis del Demo asignado y realice lo siguiente:
  - Elabore un resumen sobre: ¿En qué consiste la Demo?
  - Luego, un “Análisis Técnico” listando las tecnologías usadas, para que sirven y dando sus impresiones de cada una de ellas.
  - Por último detalle las conclusiones en las cuales deberá especificar qué fue lo que aprendió y qué le pareció la Demo asignada.