#1
Introducción al análisis de datos

Análisis y Exploración de datos | ITFS24
Prof. Martín Pasztetnik

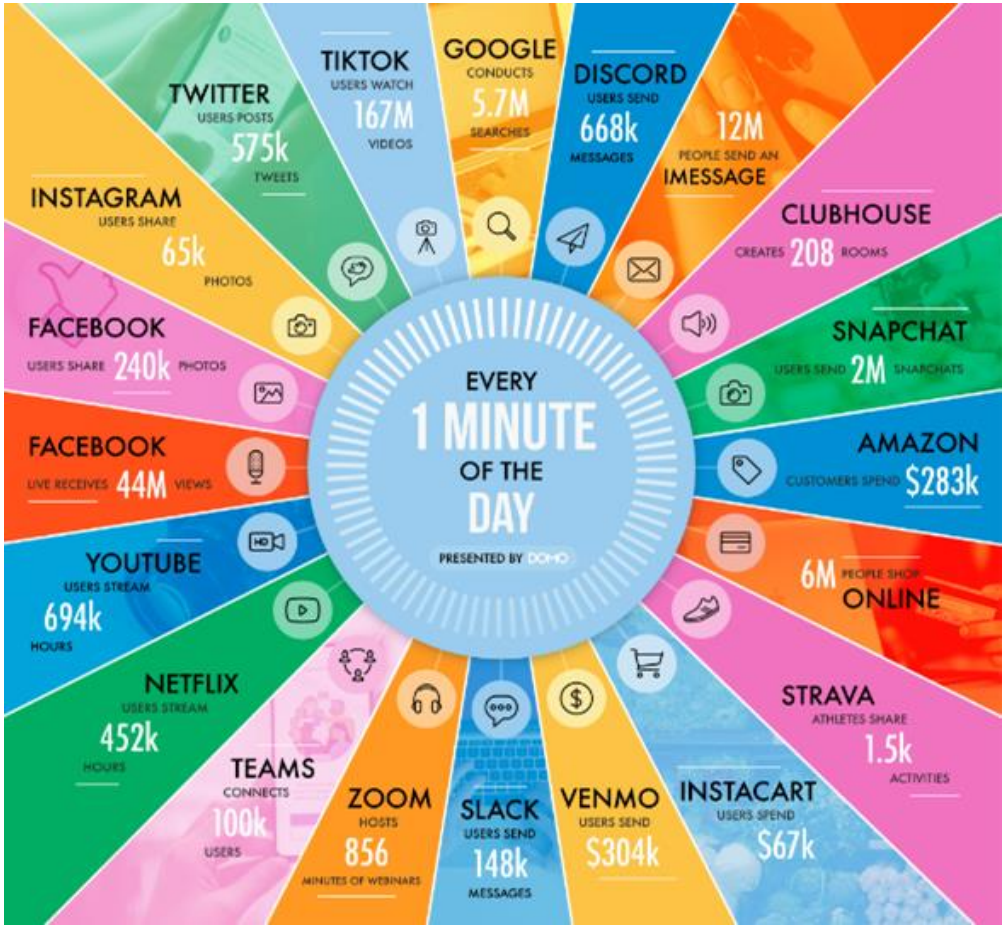
Contexto

La novedad de la ciencia de datos no radica en los últimos conocimientos científicos, sino en un cambio disruptivo en nuestra sociedad que ha sido provocado por la evolución de la tecnología: **la datificación.**

Contexto

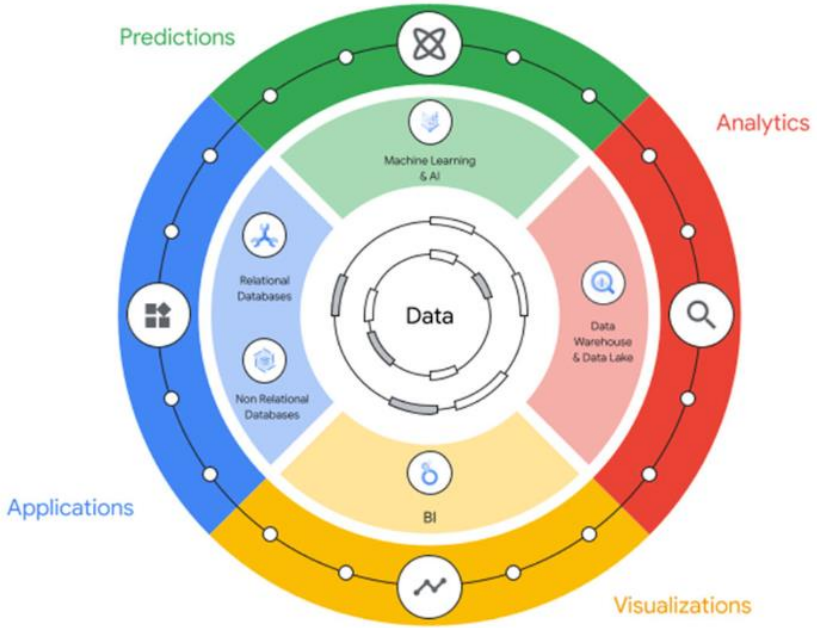
La datificación es el proceso de **convertir en datos aspectos del mundo que nunca antes habían sido cuantificados:** las redes de negocios, las listas de libros que leemos, las películas que disfrutamos, los alimentos que comemos, nuestra actividad física, nuestras compras, nuestro comportamiento al volante e incluso nuestros pensamientos (cuando los publicamos).


DATA
DATA EVERYWHERE

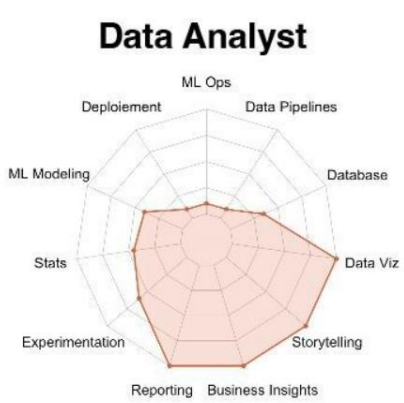

EVERY
1 MINUTE
OF THE
DAY
PRESENTED BY DOMO
TWITTER
USERS POSTS
575k
TWEETS
TIKTOK
USERS WATCH
167M
VIDEOS
GOOGLE
CONDUCTS
5.7M
SEARCHES
DISCORD
USERS SEND
668k
MESSAGES
12M
PEOPLE SEND AN
IMESSAGE
INSTAGRAM
USERS SHARE
65k
PHOTOS
CLUBHOUSE
CREATES 208 ROOMS
FACEBOOK
USERS SHARE 240k PHOTOS
SNAPCHAT
USERS SEND 2M SNAPCHATS
FACEBOOK
LIVE RECEIVES 44M VIEWS
AMAZON
CUSTOMERS SPEND $283k
YOUTUBE
USERS STREAM
694k
HOURS
6M PEOPLE SHOP
ONLINE
NETFLIX
USERS STREAM
452k
HOURS
STRAVA
ATHLETES SHARE
1.5k
ACTIVITIES
TEAMS
CONNECTS
100k
USERS
ZOOM
HOSTS
856
MINUTES OF WEBINARS
SLACK
USERS SEND
148k
MESSAGES
VENMO
USERS SEND
$304k
INSTACART
USERS SPEND
$67k

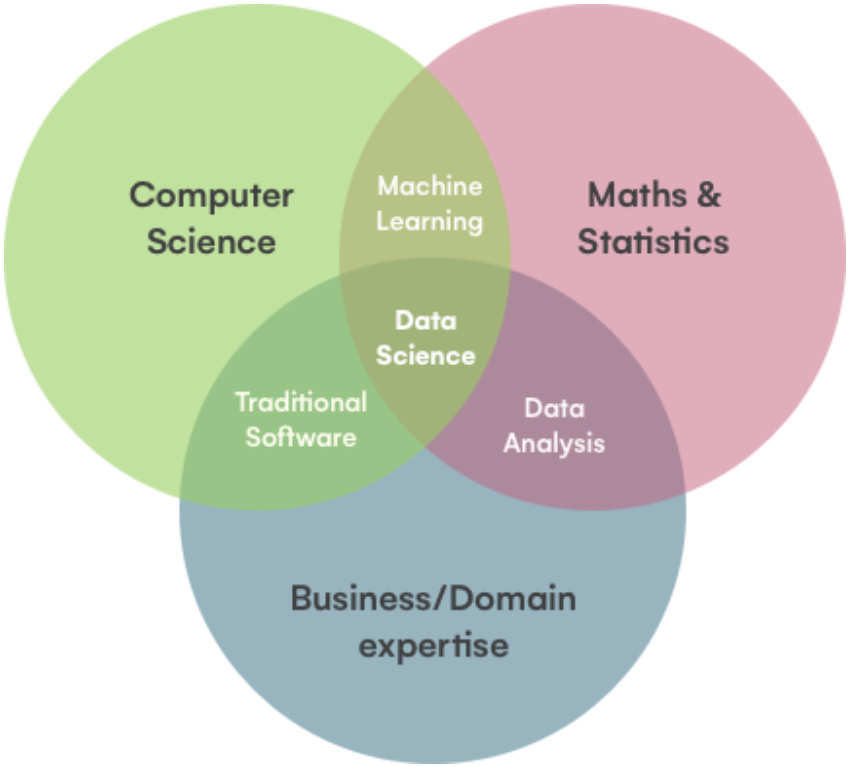
Ecosistema de datos

Los ecosistemas de datos están formados por diversos **elementos que interactúan entre sí para producir, gestionar, almacenar, organizar, analizar y compartir datos.** Estos elementos incluyen herramientas de hardware y software, y las personas que las utilizan.



Ecosistema de datos





DA

- Extracción, limpieza y calidad de los datos
- Matemáticas y estadísticas descriptiva, analizar métricas de negocio
- Extrae conclusiones relevantes basada en datos
- Excel, PowerBi, R, Python, SQL, SAS
- Herramientas de BI, Dashboards
- Habilidades en comunicación

DS

- Usa datos procesados
- Algoritmos y modelos matemáticos avanzados y algoritmos para predecir.
- Estadística probabilística, inferencial y causalidad
- Machine Learning, Deep Learning, Advanced Analytics
- Python, Flask, Docker
- Aprendizaje de las computadoras, inteligencia artificial

Especialistas en **Data Science:** tienen un gran conocimiento y dominio del manejo de datos para la **construcción de modelos complejos de información**, capaces de predecir comportamientos futuros basados en datos

Análisis de datos

Datos

Los datos son una colección de hechos que pueden ser **utilizados para sacar conclusiones, hacer predicciones y ayudar en la toma de decisiones.**

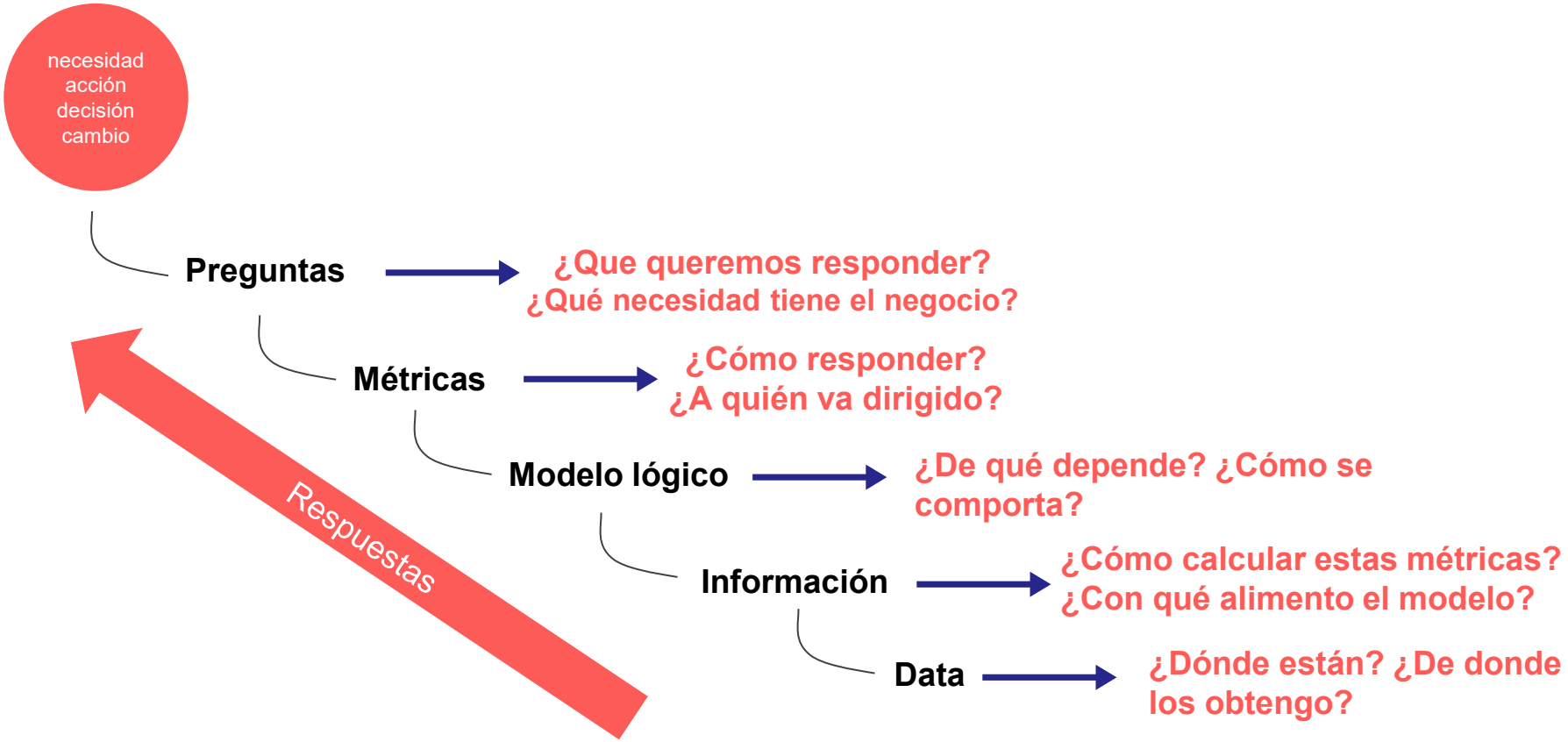
Análisis de Datos

El análisis de datos es la **recopilación, transformación y organización de datos** con el fin de sacar conclusiones, hacer predicciones y respaldar la toma de decisiones informadas.

Análisis de datos

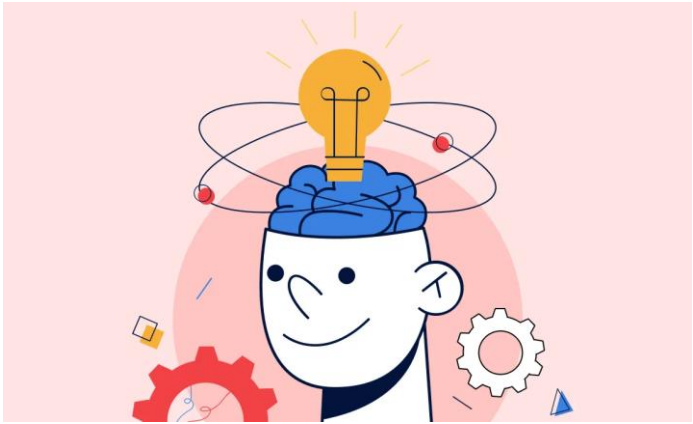
- Utilizar **hechos para guiar** la estrategia de negocio.
- Aunque el dato parezca hecho empírico, el mismo es **construido teóricamente, por lo tanto su interpretación.**
- El primer paso en la toma de decisiones basada en datos es **identificar la necesidad** del negocio.
- Los datos son más poderosos cuando se combinan con la experiencia humana, la observación y a veces la intuición.
- Es crucial incluir a expertos en la materia para analizar los resultados del análisis de datos, identificar inconsistencias y validar decisiones.

Metodología analítica



Habilidades analíticas

- Curiosidad: Deseo de aprender algo nuevo.
- **Comprensión del contexto:** Entender el entorno en el que ocurren las cosas.
- Agrupar cosas en categorías: Organizar la información para comprender el panorama general (Indicadores/Clusters)
- Mentalidad técnica: Capacidad para descomponer problemas en pasos más pequeños y trabajar con ellos de manera lógica.
- Diseño de datos: Organización de la información.



¿Cómo hacer preguntas?

Desde una perspectiva analítica...

Cuantitativas

¿Qué, Cuánto?
(Ventas, clientes, facturación, etc)

Causales

¿Por qué?

Proyectiva

Tendencia
Predicción
Cuánto voy a...

Optimización

Cómo podría...

Comparativas

¿Me va bien?
con respecto a...

Segmentación

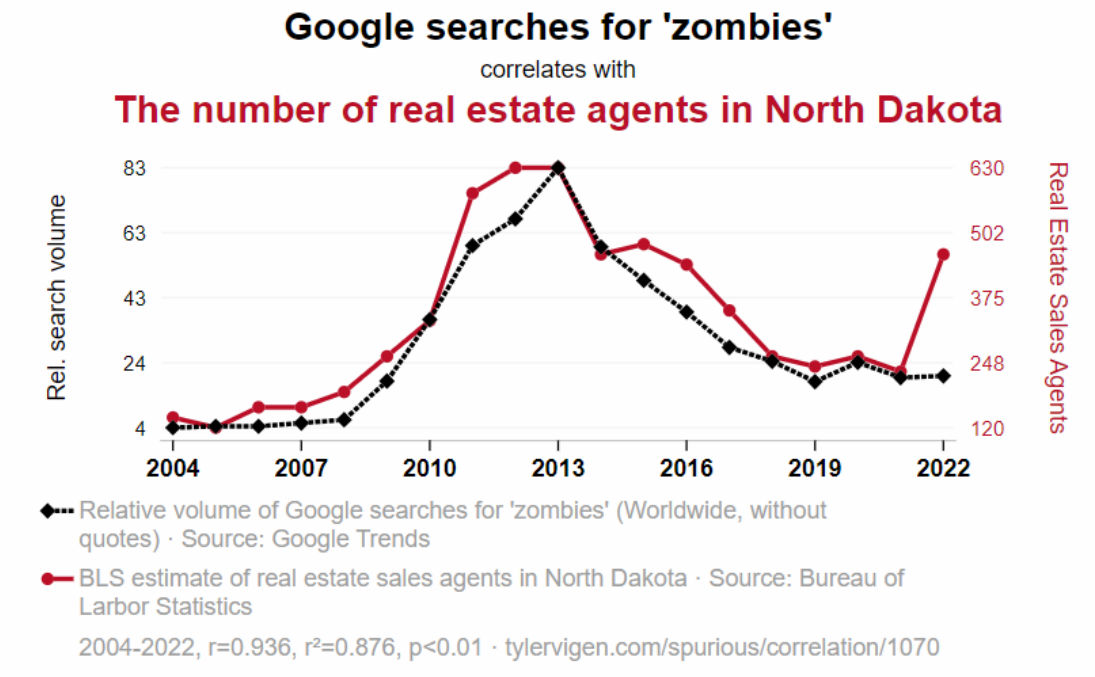
¿Quién? ¿Cuáles?

Prospectivas

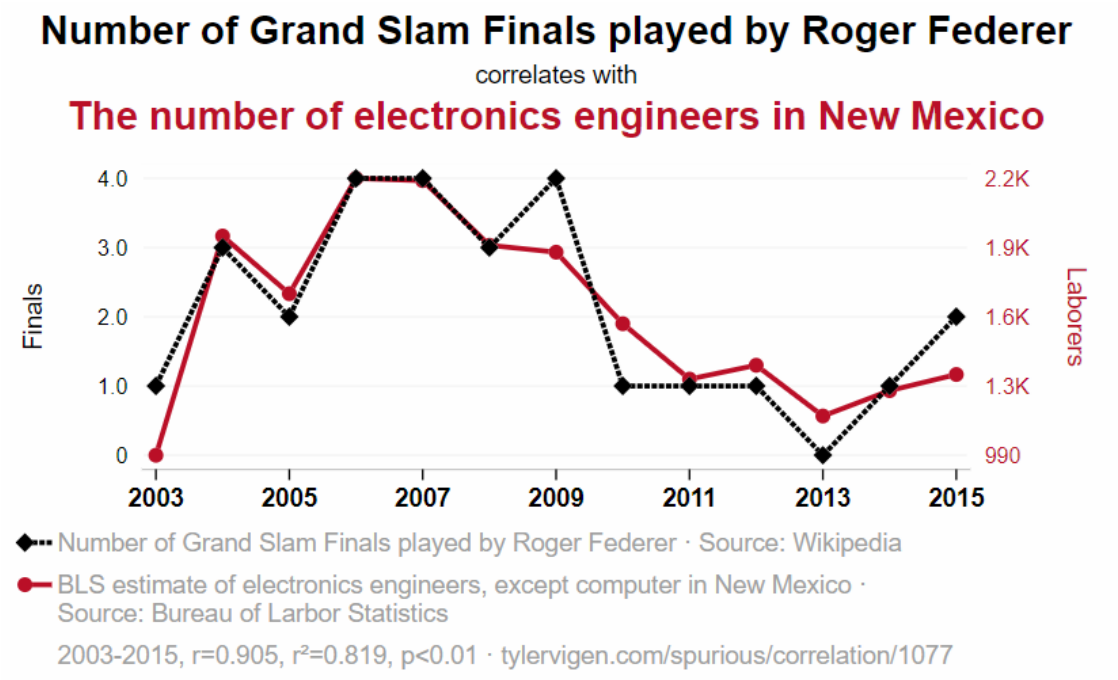
Qué sería lo ideal

Correlaciones espúreas

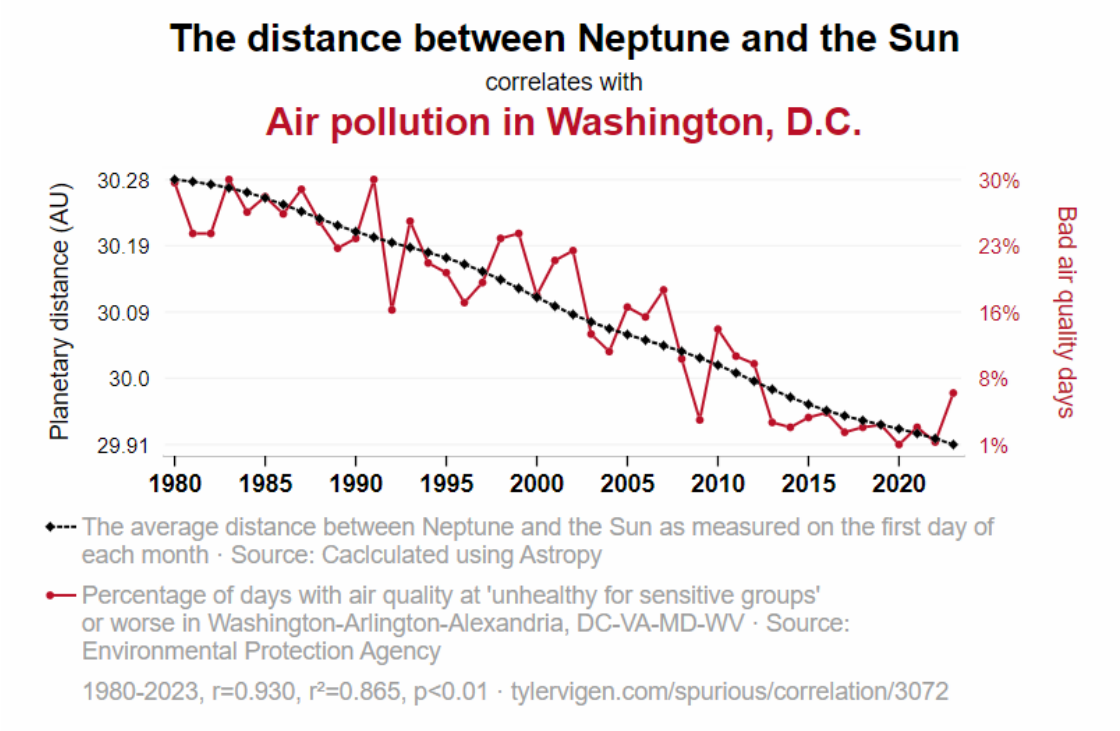
https://www.tylervigen.com/spurious-correlations



Correlaciones espúreas #2



Correlaciones espúreas #3



Especializaciones

- **Analista de marketing:** analiza las condiciones del mercado para evaluar las ventas potenciales de productos y servicios.
- **Analista de recursos humanos:** analiza los datos de nómina en busca de ineficiencias y errores.
- **Analista financiero/BI:** analiza el estado financiero recopilando, monitoreando y revisando datos.
- **Analista de riesgos:** analiza documentos financieros, condiciones económicas y datos de clientes para ayudar a las empresas a determinar el nivel de riesgo involucrado en una decisión empresarial particular.
- **Analista de salud:** analiza datos médicos para mejorar el aspecto empresarial de hospitales y centros médicos.

Proceso análisis de datos

1. EXTRAER LA DATA

Obtener datos de distintas fuentes, el volumen de los datos puede diferir (Big Data o Small Data).

→ **ETL** (Extract, Transform, Load):

Proceso de integración de datos que combina datos de múltiples fuentes de datos en un solo dataset que se carga en un sistema de destino

2. LIMPIAR Y TRANSFORMAR LA DATA

Para eliminar distorsiones y problemas de calidad. Generar nuevas variables que necesitamos para responder la pregunta.

3. ETAPA EXPLORATORIA

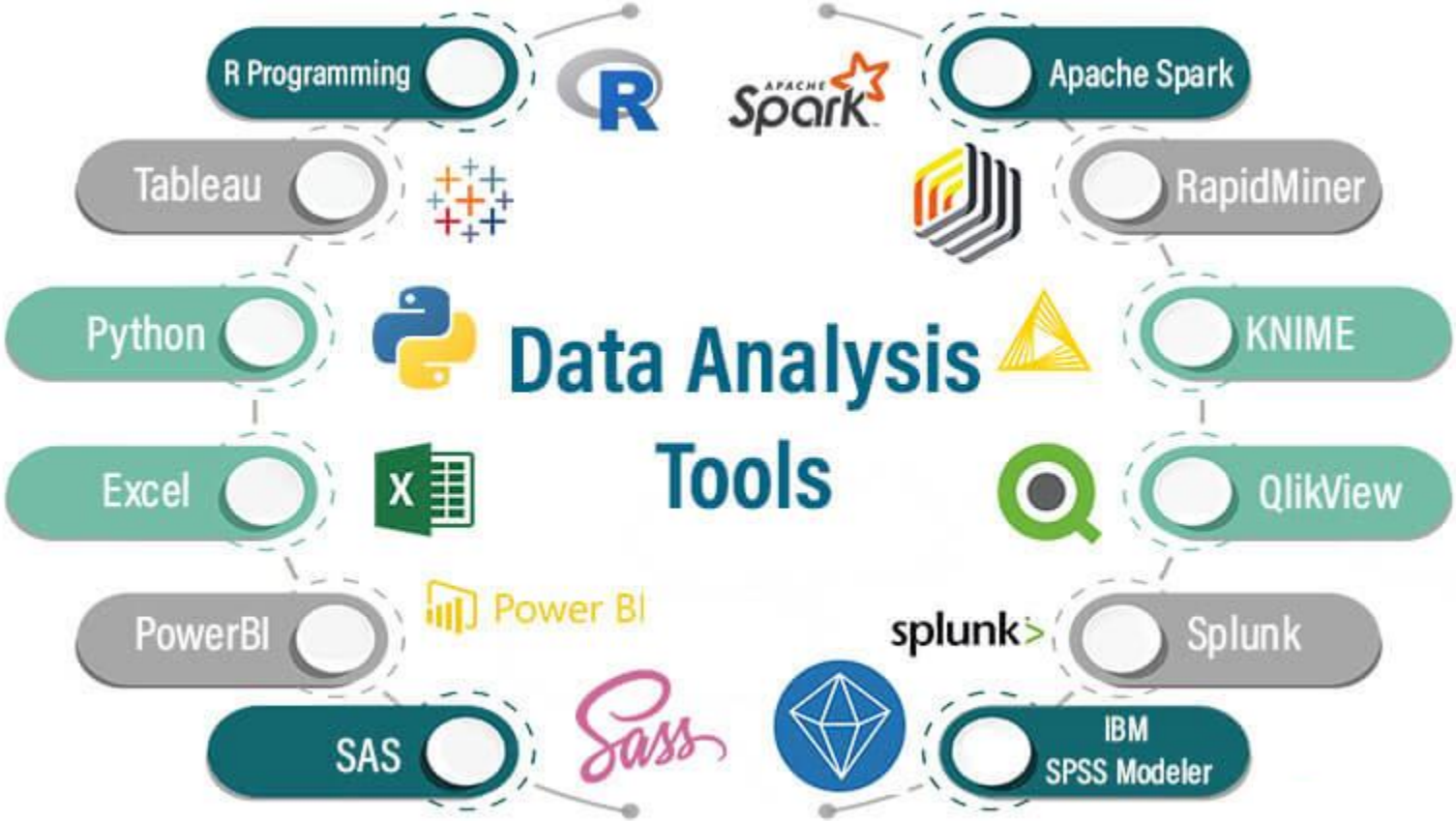
Procesar los datos usando diferentes métodos estadísticos con el objetivo de entender patrones en nuestra data.

(OPCIONAL) 4. MODELADO

Crear un modelo que pueda predecir algún comportamiento de la realidad.

5. VISUALIZAR Y NARRAR

Presentar gráficamente los datos y usar las estrategias de Data Storytelling para comunicar efectivamente los resultados.



https://datanerd.tech/

Decisiones basadas en datos

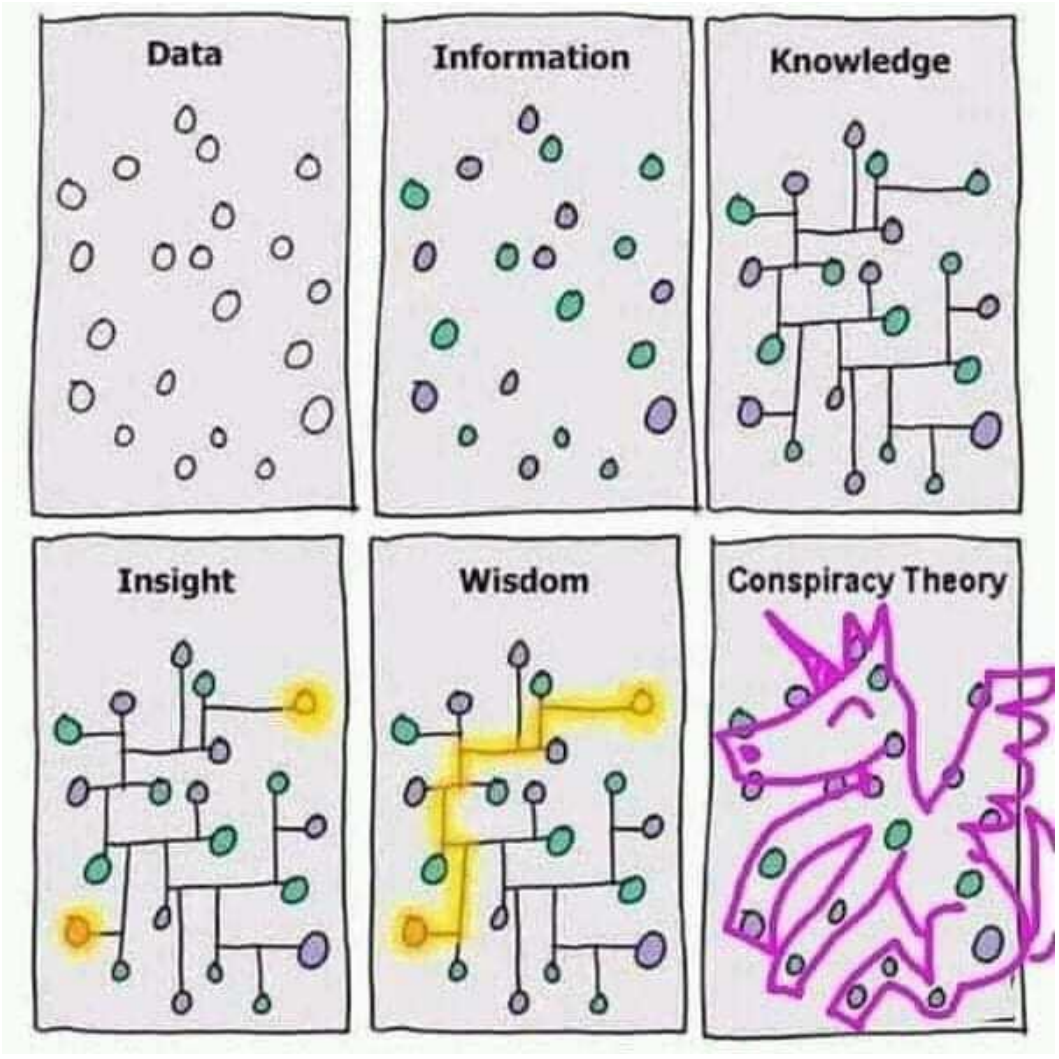
- **Hacer predicciones:** Utilizar datos para tomar decisiones informadas sobre cómo podrían ser las cosas en el futuro.
- **Categorizar cosas:** Asignar información a diferentes grupos o clústeres basados en características comunes.
- **Detectar algo inusual:** Identificar datos que difieren de la norma.
- **Identificar temas:** Llevar la categorización un paso más allá al agrupar la información en conceptos más amplios.
- **Descubrir conexiones:** Encontrar desafíos similares enfrentados por diferentes entidades, y luego combinar datos y conocimientos para abordarlos.
- **Encontrar patrones:** Utilizar datos históricos para comprender lo que sucedió en el pasado y, por lo tanto, es probable que vuelva a ocurrir.

S M A R T

				
S-pecific Is the question specific? Does it address the problem? Does it have context? Will it uncover a lot of the information you need?	**M-easurable** Will the question give you answers that you can measure?	**A-ction-oriented** Will the answers provide information that helps you devise some type of action plan?	**R-elevant** Is the question about the particular problem you are trying to solve?	**T-ime-bound** Are the answers relevant to the specific time being studied?

Toma de decisiones basada en datos

- El objetivo de los analistas de datos: **llegar a conclusiones precisas y hacer recomendaciones acertadas.**
- Todo comienza con **datos completos, correctos y relevantes.**
- **Interpretación** precisa de los datos: crucial para evitar pérdidas y tomar decisiones acertadas.
- **Uso estratégico de los datos:** puede transformar y aumentar los ingresos de las empresas.
- Diferencia entre datos incompletos y una cantidad pequeña de datos: el riesgo de decisiones erróneas frente a la posibilidad de tomar decisiones acertadas con pruebas pequeñas y precisas.


Data
Information
Knowledge
Insight
Wisdom
Conspiracy Theory

Tipos de datos

- **Datos cuantitativos**

Los datos cuantitativos se refieren a medidas específicas y objetivas de hechos numéricos. Esto a menudo puede ser el qué, cuántos y con qué frecuencia sobre un problema. En otras palabras, cosas que se pueden medir, como un **número, cantidad o rango.**

- **Datos cualitativos**

Los datos cualitativos son una **medida subjetiva** y explicativa de una cualidad o característica. Básicamente, son las cosas que no se pueden medir con datos numéricos, como el color de tu cabello. Los datos cualitativos son excelentes para ayudarnos a responder preguntas de por qué.


QUANTITATIVE
DATA
NUMERICAL
DISCRETE
COUNTING
CONTINUOUS
MEASUREMENT
FACTUAL
QUALITATIVE
DATA
DESCRIPTIVE
SENSES
SEE
FEEL
SMELL
HEAR
TASTE
SUBJECTIVE

Tipos de datos

- **Datos estructurados**

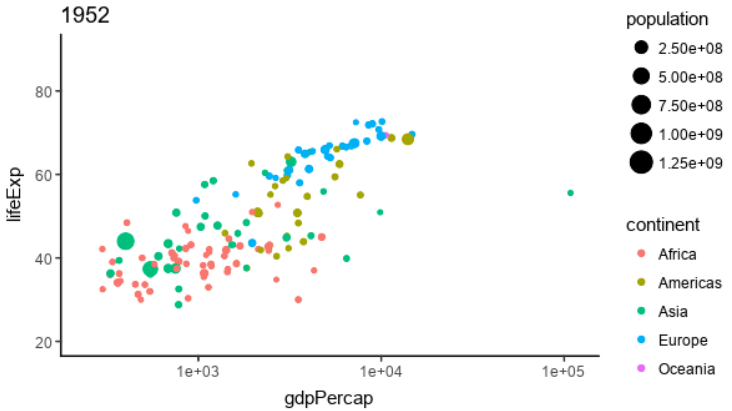
Los datos organizados en un formato específico, como filas y columnas. Esto facilita el almacenamiento y la consulta para las necesidades comerciales. Si los datos se exportan, la estructura se mantiene junto con los datos.

- **Datos no estructurados**

Datos que no pueden almacenarse como columnas y filas en una base de datos relacional. Estos son datos que no están organizados de ninguna manera fácilmente identificable. Los archivos de audio y video son ejemplos de datos no estructurados porque no hay una forma clara de identificar u organizar su contenido.

Visualización de datos

- Ayudan a dar cuenta de **comportamientos** que no eran tan obvios
- Los **patrones** quedan a la vista con mayor facilidad
- Permite **identificar errores y datos faltantes** en datasets durante el análisis
- Las métricas pueden ser visualizadas de manera **más efectiva**
- Los tableros bien elaborados, permiten comunicar de manera universal



Sesgo en datos

- Una preferencia consciente o inconsciente a favor o en contra de una persona, grupo de personas o cosa.
- **Prejuicio de datos:** Cuando una preferencia a favor o en contra de una persona, grupo de personas o cosa sesga sistemáticamente los resultados del análisis de datos en una dirección determinada.
- **Reconociendo el prejuicio:** Una vez que conocemos y aceptamos que tenemos prejuicios, podemos comenzar a reconocer nuestros propios patrones de pensamiento y aprender a manejarlos.
- Tipos de prejuicio:

  Prejuicio de muestreo: Sobre o subrepresentación de ciertos miembros de una población como resultado de trabajar con una muestra que no es representativa de la población en su conjunto.

  Prejuicio del experimentador, observador o de investigación.

Credibilidad

Buena fuente de datos: Una fuente de datos que es confiable, original, completa, actual y citada.

Mala fuente de datos: Una fuente de datos que no es confiable, original, completa, actual y citada.

Calidad de los datos

- El **origen** de los datos es un factor crítico a la hora de decidir qué usar.
- El uso de datos no confiables o no validados, pueden generar conclusiones erróneas, discontinuidad, no escalabilidad, etc.
- **Ahorramos tiempo y recursos** en los procesos y en el uso de sistemas de información
- Hace que los análisis sean confiables; **toma de decisiones oportunas**

Dimensiones de la calidad del dato #1

Conformidad

La **procedencia** y la **trazabilidad** del dato son características que hacen a la fiabilidad.

Si trabajamos con una tabla de la que no conocemos su procedencia o bien que bajamos de una página poco confiable, no tendrán conformidad y nos pueden llevar a análisis erróneos.

Asimismo, si no podemos reconstruir el camino completo del dato desde su captura hasta la actualidad, el set de datos no tiene trazabilidad.

Actualización

Los datos deben estar actualizados.

Un dataset sin referencias de la fecha de confección o de la fecha de último update puede distorsionar el análisis y no permite interpretaciones temporales

Integridad

Los datos deben ser accesibles con bajo nivel de esfuerzo.

Deben estar en lugares previsibles y ser fácilmente ubicables y legibles.

Ejemplo 1: una tabla con nombre de campos numerados (Campo1, Campo2..., etc).

Ejemplo 2: Una tabla que se aloja en un directorio poco habitual

Dimensiones de la calidad del dato #2

Completitud

Los datos deben estar completos.

Tablas donde un atributo importante tiene valores nulos/datos perdidos es una de las situaciones más comunes en la ciencia de datos.

Ejemplo: tabla con datos filiatorios y de contactación con campos vacíos.

Validez

Los datos deben tener exactitud y validez, no deben ser ambiguos.

Se debe tener la certeza de que son datos reales.

Por otro lado, deben tener formatos y tipologías válidas. **Unicidad**: que no haya datos duplicados

Consistencia

Interna: Calidad de caracteres y de lo que se guarda en los campos.

Externa: Calidad de interdependencia y relacionabilidad de los campos.

Ejemplo: Las Primary Keys y las Foreign Keys deben ser consistentes y permitir la relación entre tablas.

Ética

Ética de datos: Estándares bien fundamentados de lo correcto y lo incorrecto que dictan cómo se recopilan, comparten y utilizan los datos.

Aspectos de la ética de datos:

- Consentimiento.
- Privacidad de datos.
- Transparencia de la transacción.
- Apertura.
- Propiedad.

Anonimización de datos: El proceso de proteger los datos privados o sensibles de las personas mediante la eliminación de información de identificación.

Interoperabilidad de datos: Factor clave que conduce al uso exitoso de datos abiertos entre empresas y gobiernos. Es la capacidad de sistemas y servicios de datos para conectarse y compartir datos abiertamente.

Gobernanza de datos

La gobernanza de datos se refiere al conjunto de procesos, políticas y procedimientos que garantizan la **disponibilidad, integridad, calidad y seguridad de los datos en una organización**. Es fundamental para garantizar el uso adecuado y eficiente de los datos en todas las áreas de la empresa. Algunos aspectos clave de la gobernanza de datos incluyen:

- Establecimiento de políticas y procedimientos para la gestión de datos.
- Definición de roles y responsabilidades para la administración de datos.
- Garantía de la calidad y consistencia de los datos.
- Protección de la privacidad y seguridad de los datos.
- Cumplimiento de regulaciones y normativas relacionadas con los datos.

Repositorio datasets

https://www.kaggle.com/

https://datasetsearch.research.google.com/

https://archive.ics.uci.edu/

https://opendatacommons.org/

https://www.indec.gob.ar/

https://www.estadisticaciudad.gob.ar/eyc/

https://www.datos.gob.ar/