

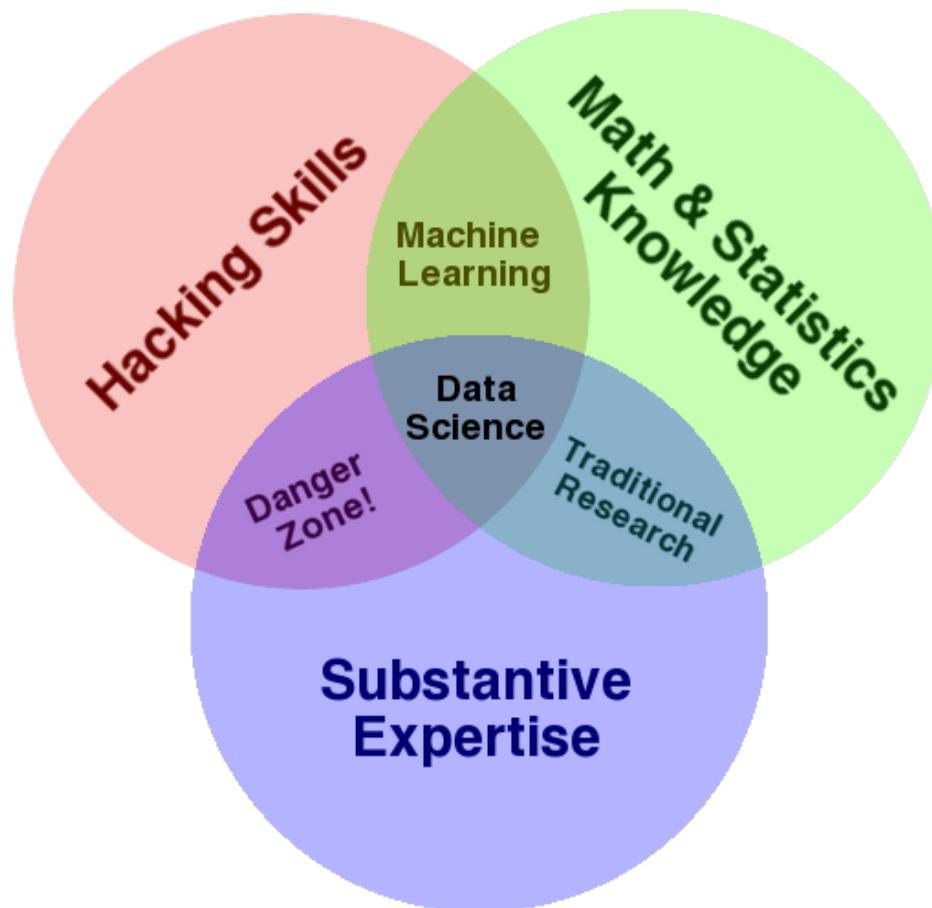
# Ciencia de datos

## “Desde cero”

# ¿Qué es la Ciencia de datos?

- Obtener insights accionables a partir de la información para tomar decisiones más inteligentes.
- El deseo de ir más allá de la superficie de un problema por medio de datos, hacer las preguntas correctas, derivar una serie de hipótesis y probarlas.

# ¿Qué es la Ciencia de datos?



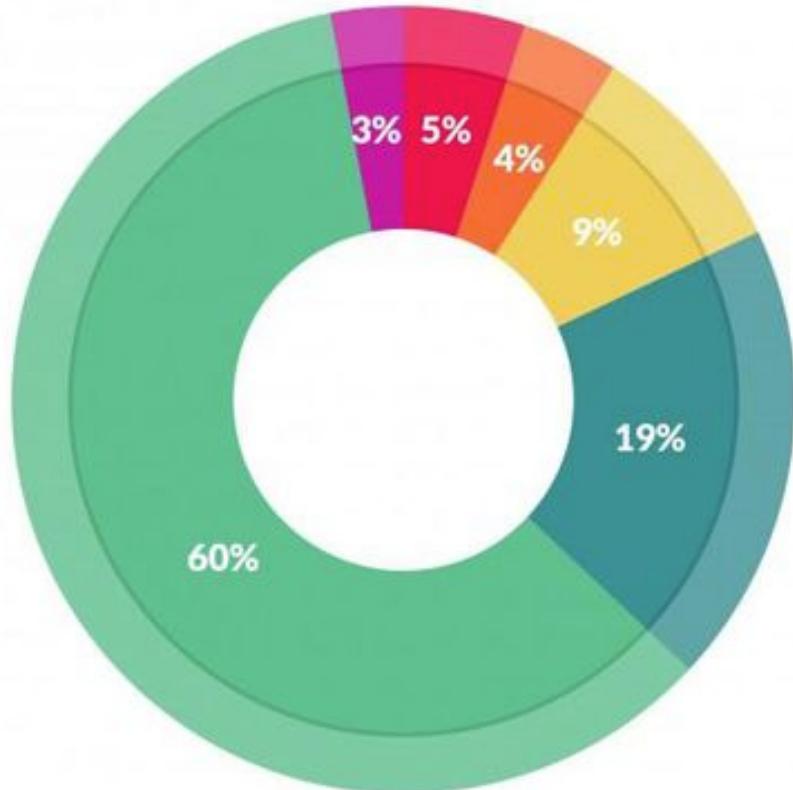
Fuente: <http://drewconway.com/>

# ¿Qué es la Ciencia de datos?

“El trabajo más sexy del siglo 21”

Realidad: **Tedioso, mucho tiempo es invertido en  
limpieza de datos.**

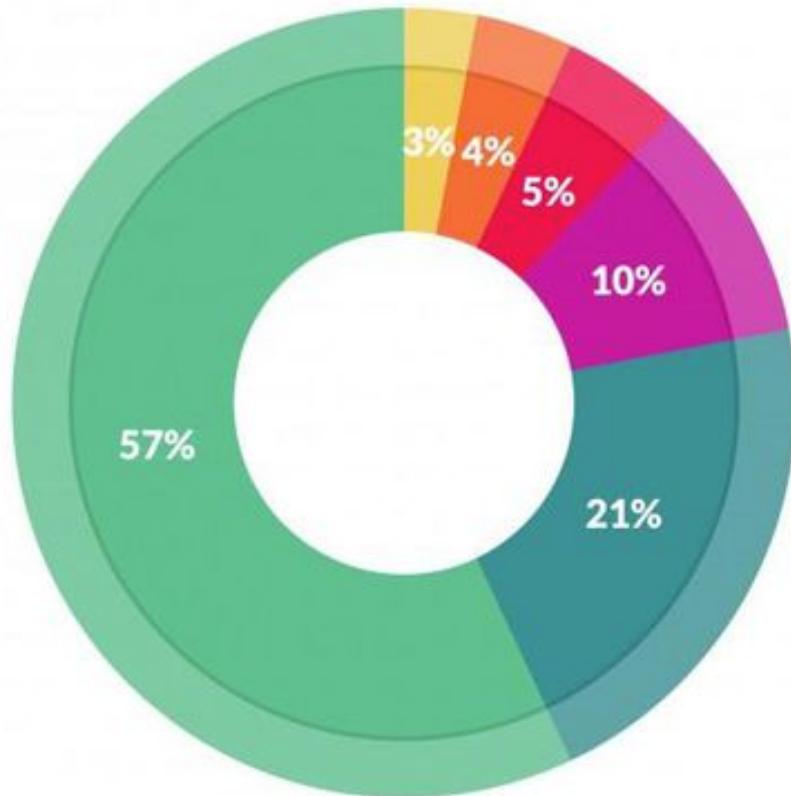
# ¿Qué es la Ciencia de datos?



What data scientists spend the most time doing

- *Building training sets:* 3%
- *Cleaning and organizing data:* 60%
- *Collecting data sets:* 19%
- *Mining data for patterns:* 9%
- *Refining algorithms:* 4%
- *Other:* 5%

# ¿Qué es la Ciencia de datos?

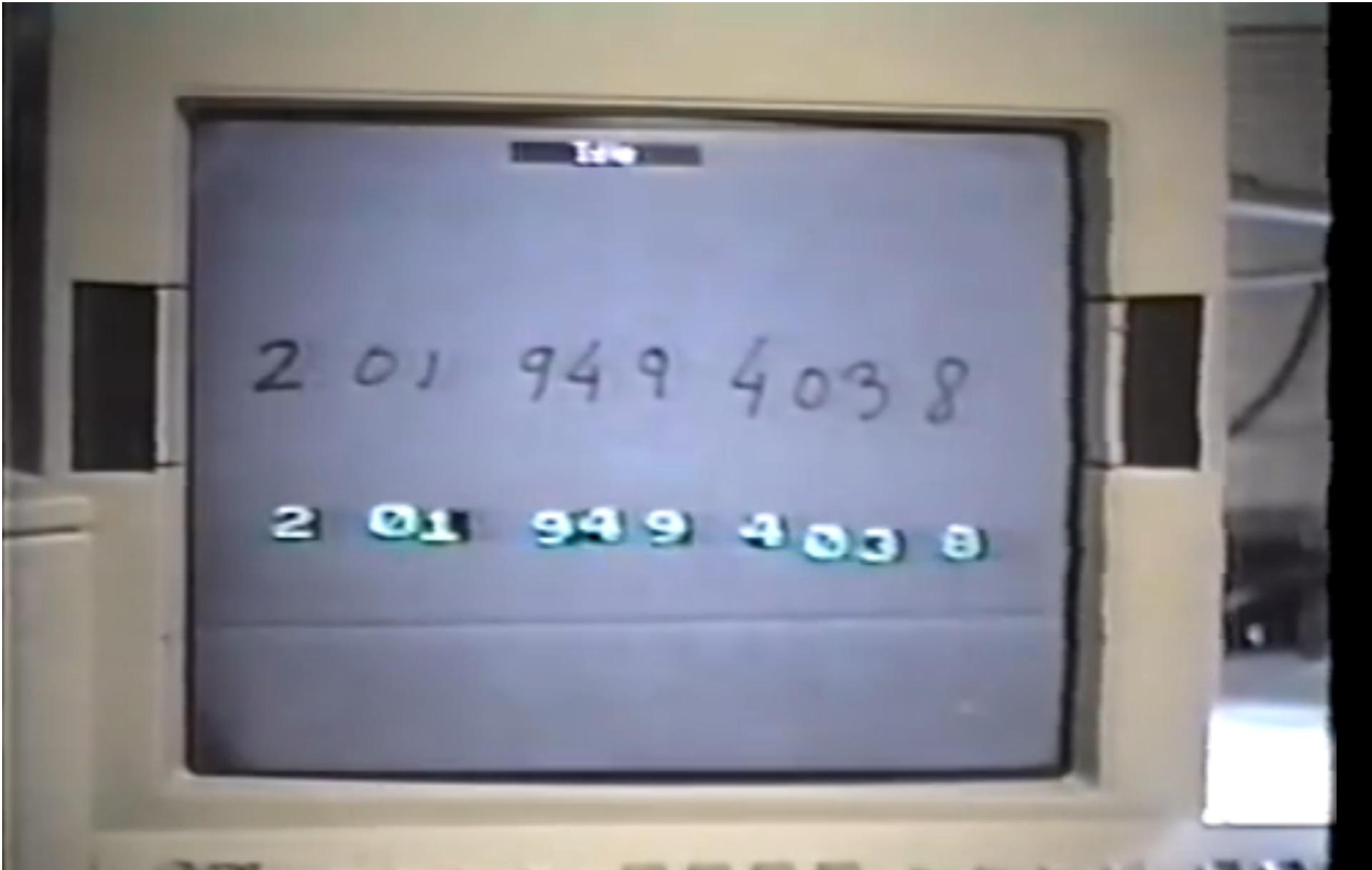


What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

# ¿Es algo nuevo?

- ¡No!
- Siempre ha existido, simplemente se le dió el nombre recientemente englobando diferentes áreas.
  - Análisis del negocio
  - Estadística
  - Ciencias Computacionales
  - Análisis de Datos.



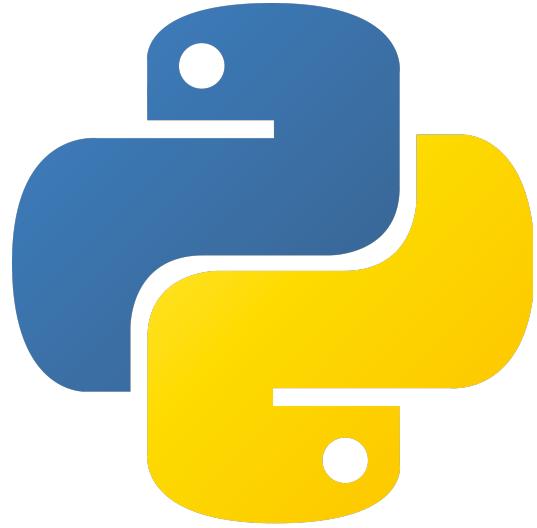
Reconocimiento de dígitos escritos a mano con una red neuronal convolucional (1993)

# Nos estamos ahogando en datos

- Páginas web
- Videos
- Imágenes
- Smartphone
- Smartwatch
- Wearables
- IoT



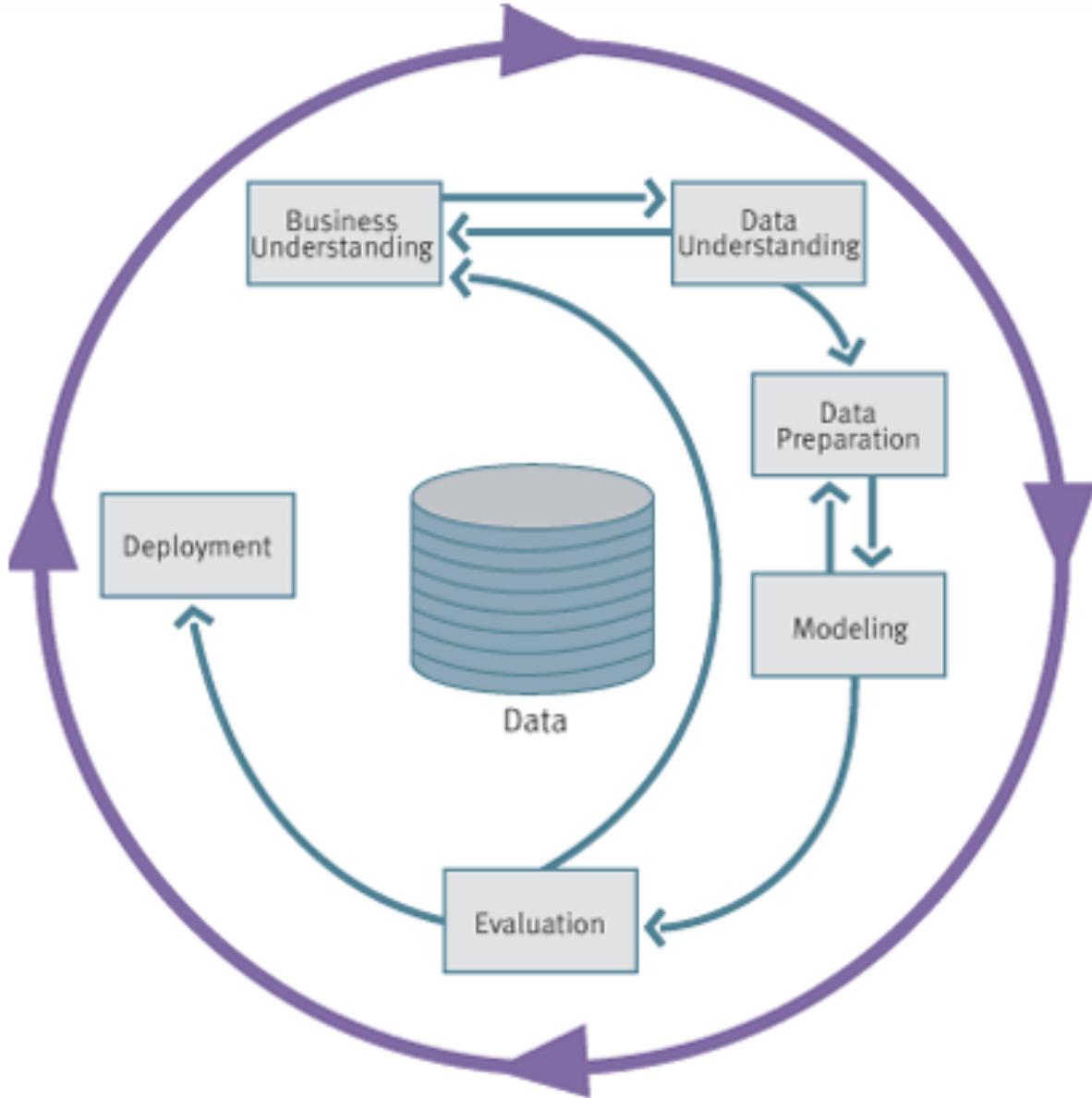
VS



- Hecho por y para estadísticos (no sea el código más eficiente)
- Análisis Exploratorio
- Lenguaje de propósito general.
- Comunidad más grande y diversa.
- Integración web sencilla

# CRISP DM Methodology

Cross-industry standard process for data mining



## Business Understanding

### **Compreensión del negocio (Objetivos y requerimientos desde una perspectiva no técnica)**

- Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)
- Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio,...)
- Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito)
- Generación del plan del proyecto (plan, herramientas, equipo y técnicas)

## Data Understanding

### **Compreensión de los datos** (Familiarizarse con los datos teniendo presente los objetivos del negocio)

- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos
- Verificación de calidad de datos

## Data Preparation

# **Preparación de los datos** (Obtener la vista minable o dataset)

- Selección de los datos
- Limpieza de datos
- Construcción de datos
- Integración de datos
- Formateo de datos

## Modeling

**Modelado** (Aplicar las técnicas de minería de datos a los dataset)

- Selección de la técnica de modelado
- Diseño de la evaluación
- Construcción del modelo
- Evaluación del modelo

## Evaluation

**Evaluación** (De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio)

- Evaluación de resultados
- Revisar el proceso
- Establecimiento de los siguientes pasos o acciones

## Deployment

**Despliegue** (Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización)

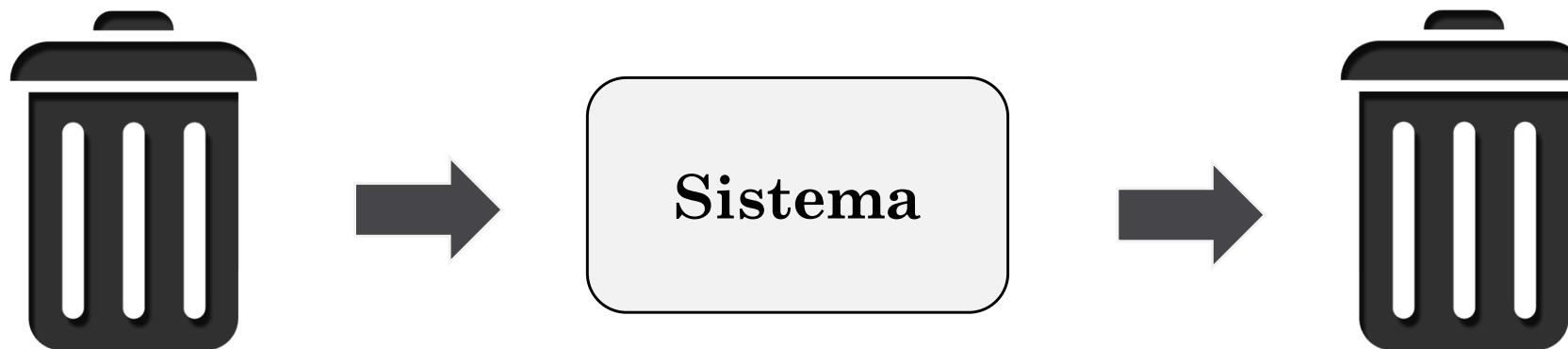
- Planificación de despliegue
- Planificación de la monitorización y del mantenimiento
- Generación de informe final
- Revisión del proyecto

# ¿Podemos predecir lo impredecible?



# Datos

- Los datos / información que utilicemos puede no ser perfecta.
- El modelo es tan bueno como la información que se utiliza.



# Manejo de datos faltantes “missing values”

- Omitir datos (borrar)
- Imputación de datos (llenar)
  - Sustituir con la mediana/media de los datos
  - Interpolación
  - kNN (Valores de vecinos más cercanos)

# Expresiones regulares

- Muy utilizadas para encontrar patrones en texto y limpiar datos según sea conveniente.
- Ej. Extraer todos los dígitos de un string
  - “Av. Eugenio Garza Sada 2501” → “2501”

# Modelos

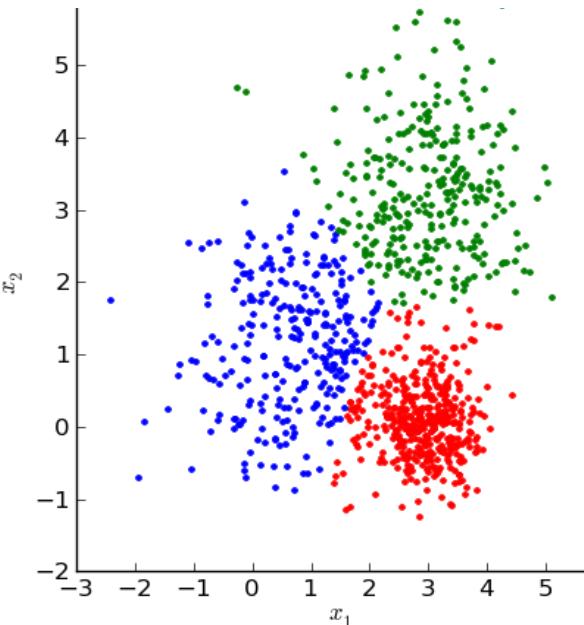
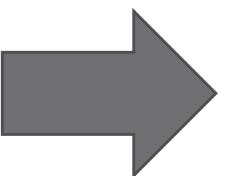
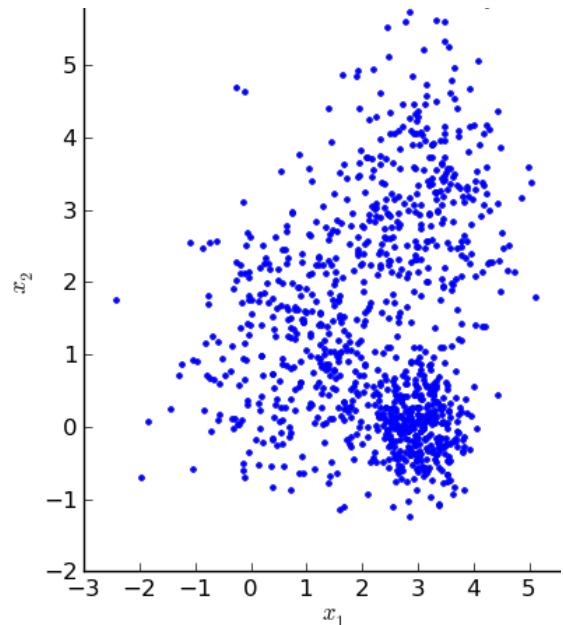
- Un modelo es una **representación** física, matemática o lógica de un proceso, sistema, entidad o fenómeno.
- Puede ser considerado como una aproximación a la realidad.
- No podemos modelar cada aspecto.

# Clasificación y Regresión

- Clasificación
  - La intención es predecir para cada individuo de una población a qué clase pertenece (usualmente mutuamente excluyentes).
  - Ej. Spam vs No spam, si una persona le dará click a una ad o no, si el día de mañana estará soleado, nublado, etc.
- Regresión
  - La intención es predecir para cada individuo de una población un valor numérico de una variable.
  - Ej. El costo de una casa, qué tanto va a utilizar un cliente cierto servicio, el valor del Bitcoin el día de mañana.
- **En resumen:**
  - Regresión para valores continuos, clasificación para valores discretos

# Clustering

- Su intención es agrupar individuos de una población mediante similaridades o patrones en común.
- Ej. Diferentes Estilos de vida de las personas.



# DATA FALLACIES TO AVOID



## CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.



## COBRA EFFECT

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.



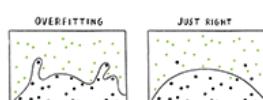
## SAMPLING BIAS

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.



## REGRESSION FALLACY

When something happens that's unusually good or bad, it will revert back towards the average over time.



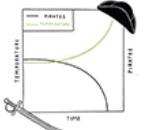
## OVERFITTING

Creating a model that's overly tailored to the data you have and not representative of the general trend.



## DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.



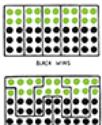
## FALSE CAUSALITY

Falsely assuming when two events appear related that one must have caused the other.



## SURVIVORSHIP BIAS

Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.



## GERRYMANDERING

Manipulating the geographical boundaries used to group data in order to change the result.



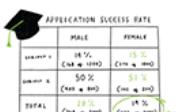
## GAMBLER'S FALLACY

Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).



## HAWTHORNE EFFECT

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.



## SIMPSON'S PARADOX

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



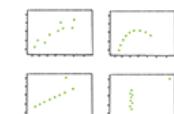
## MCNAMARA FALLACY

Relying solely on metrics in complex situations and losing sight of the bigger picture.



## PUBLICATION BIAS

Interesting research findings are more likely to be published, distorting our impression of reality.



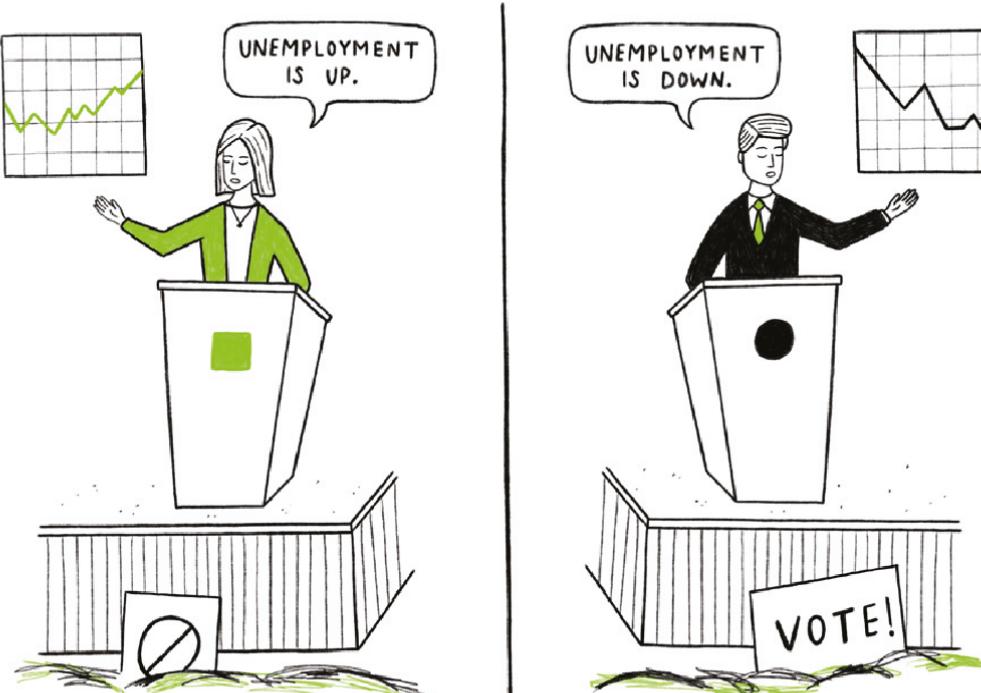
## DANGER OF SUMMARY METRICS

Only looking at summary metrics and missing big differences in the raw data.



# Cherry Picking

- Seleccionar resultados que respaldan tu hipótesis, excluyendo aquellos que no.



# Data Dredging

- No percatarse de que la correlación encontrada fue cuestión de azar.



# Survivorship Bias

- Derivar conclusiones de un set incompleto de datos, porque esos datos sobrevivieron algún criterio de selección.



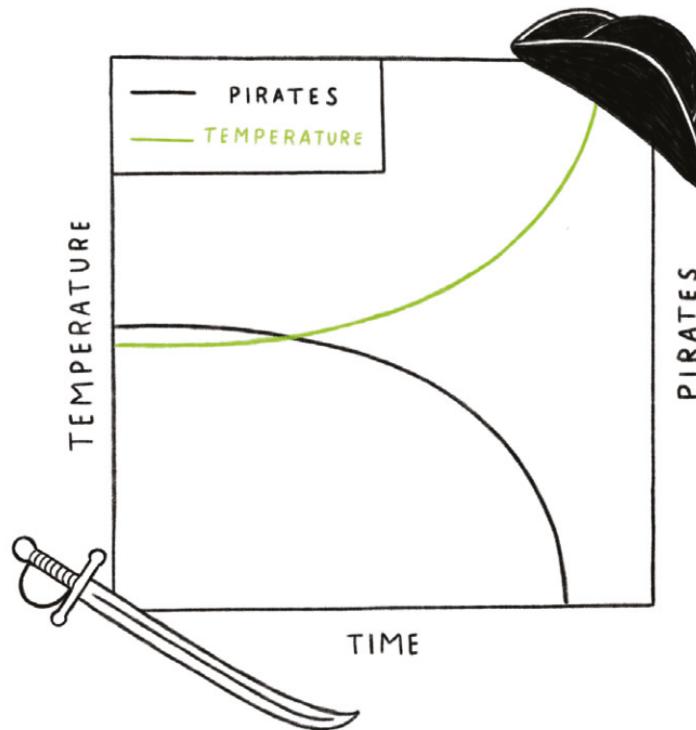
# Cobra Effect

- Cuando una iniciativa produce el resultado opuesto al previsto.



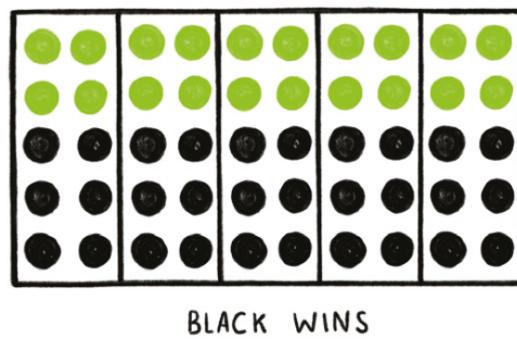
# False Causality

- Erróneamente suponer que cuando dos eventos ocurren juntos, uno es consecuencia del otro.



# Gerrymandering

- La práctica de manipular deliberadamente las fronteras para comunicar resultados que favorezcan.



# Sampling Bias

- Derivar conclusiones de un set de datos que no es representativo de la población.



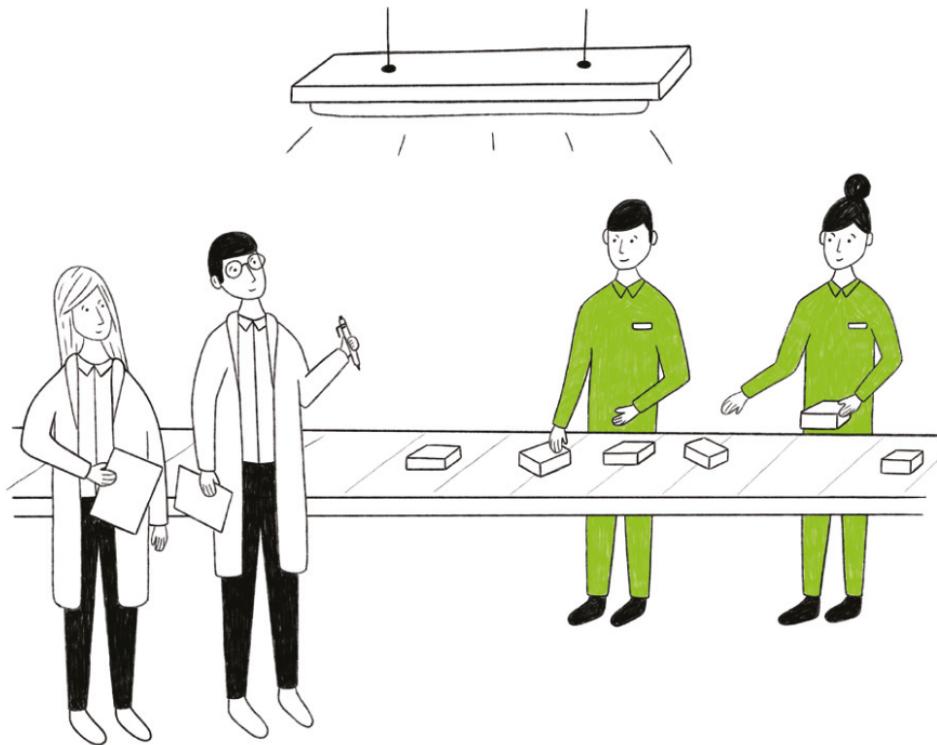
# Gambler's Fallacy

- La creencia errónea de que porque algo ocurrió con más frecuencia de lo usual, es menos probable que ocurra en un futuro.



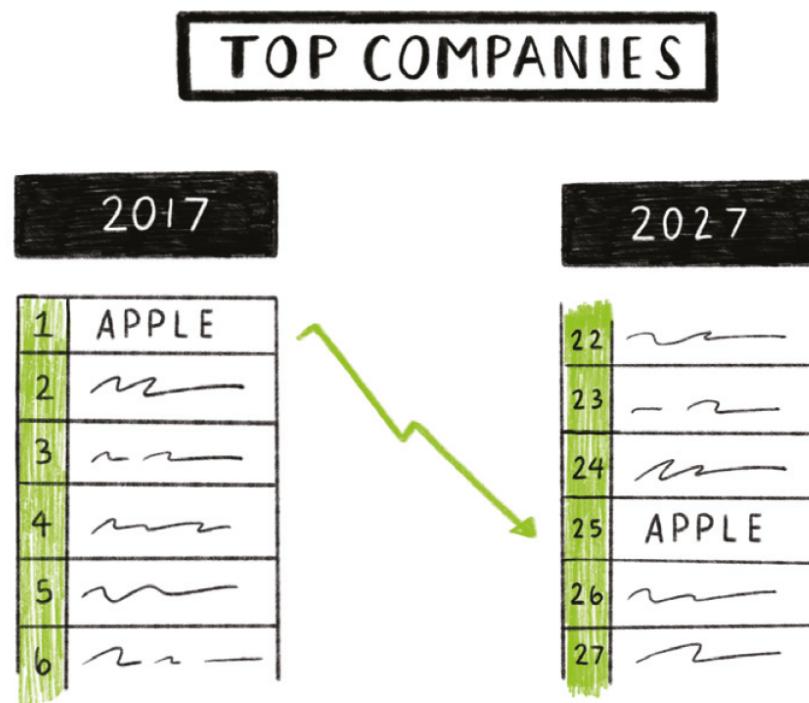
# Hawthorne Effect

- Cuanto el acto de monitorear afecta los resultados.  
También conocido como el efecto observador.



# Regression Toward the Mean

- Cuando algo ocurre inusualmente bueno o malo, en algún momento se revertirá al promedio.



# Simpson's Paradox

- Cuando una tendencia aparece en diferentes grupos de datos, pero desaparece o se revierte cuando los grupos se combinan.

APPLICATION SUCCESS RATE

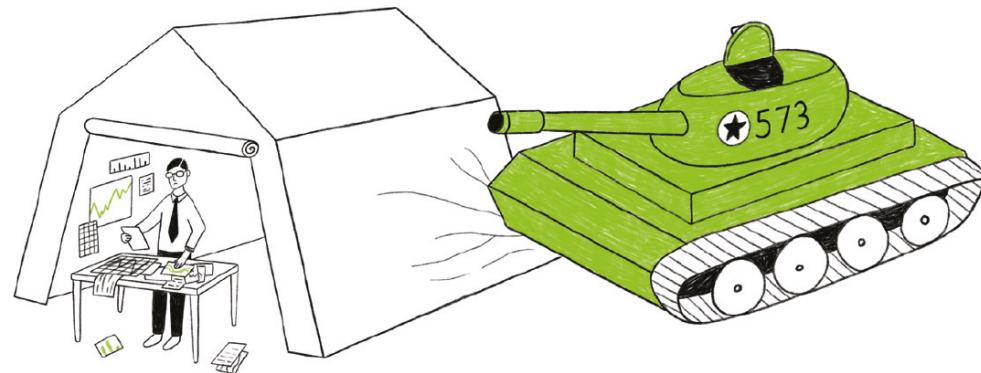


	MALE	FEMALE
SUBJECT 1	14 % (168 of 1200)	15 % (270 of 1800)
SUBJECT 2	50 % (400 of 800)	51 % (102 of 200)
TOTAL	28 % (568 of 2000)	19 % (372 of 2000)

??

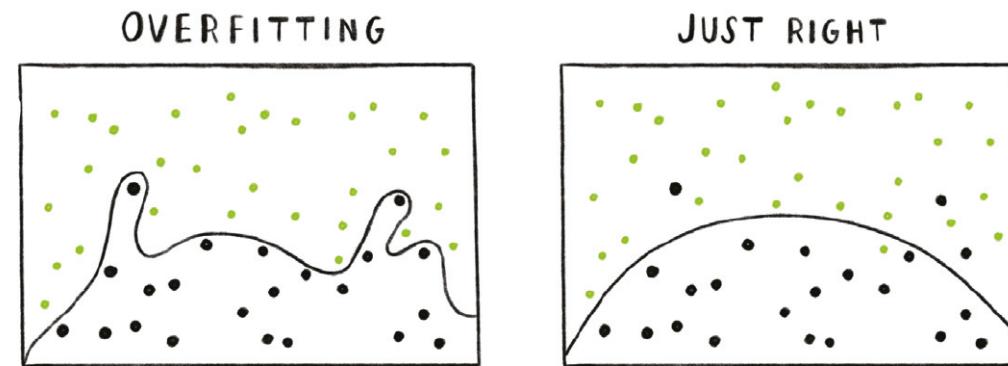
# McNamara Fallacy

- Cuando se confía únicamente en las métricas en situaciones complejas se puede perder la perspectiva más amplia.



# Overfitting

- Una explicación más compleja se ajustará más a los datos que una más sencilla. Sin embargo, comúnmente es más representativa ésta última.



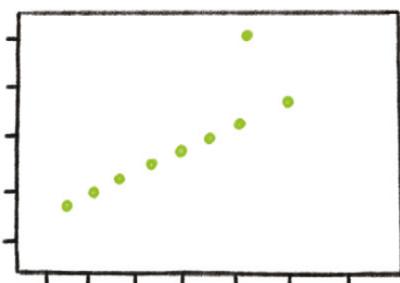
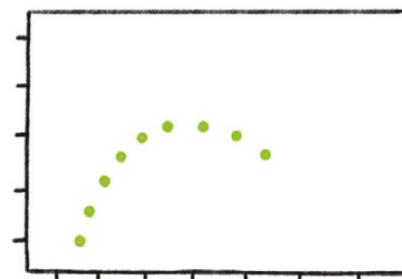
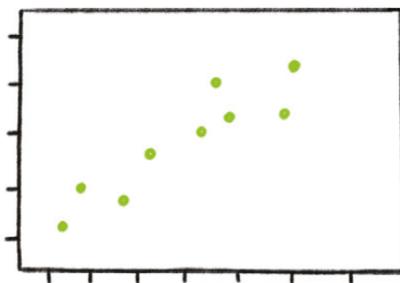
# Publication Bias

- Entre más interesante sea el hallazgo de investigación es más probable que sea publicado.



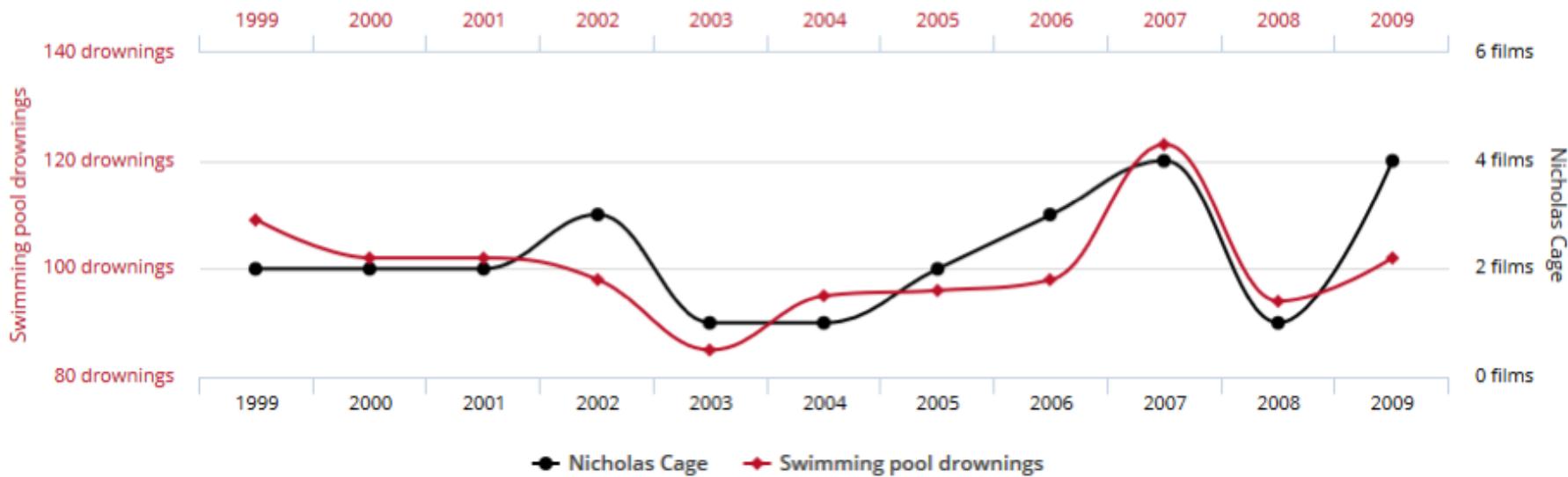
# Danger of Summary Metrics

- Puede ser confuso analizar únicamente los promedios de las métricas.



# Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

Correlation: 66.6% ( $r=0.666004$ ,  $p>0.05$ )

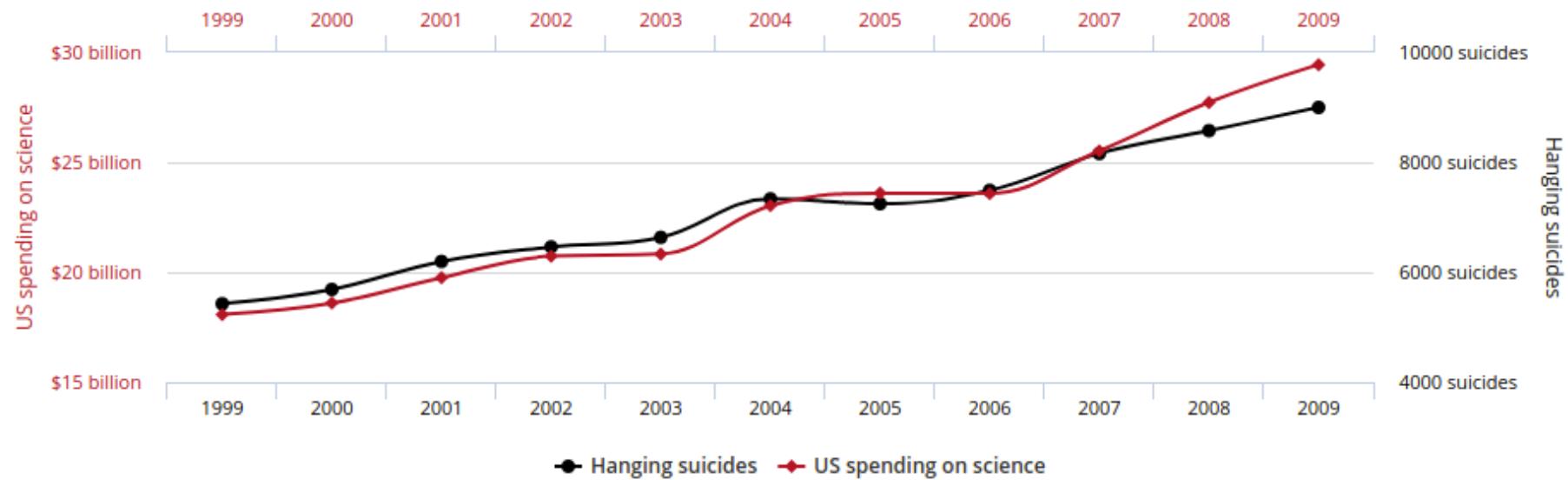


Data sources: Centers for Disease Control & Prevention and Internet Movie Database

[tylervigen.com](http://tylervigen.com)

# US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation

Correlation: 99.79% ( $r=0.99789126$ )



Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

[tylervigen.com](http://tylervigen.com)

# Librerías de Python para DS

- **Numpy** - álgebra lineal, operaciones de arreglos n-dimensionales.
- **Pandas** - Operaciones con bases de datos estructuradas, muy utilizada para pre-procesamiento de datos y análisis exploratorio.
- **Scikit Learn** – Machine learning (clasificación, regresión, clustering, y reducción de dimensionalidad).
- **Matplotlib** – Visualizaciones de gráficas (histogramas, gráficas de barras, lineales, etc.).
- **Seaborn** – Más visualizaciones.
- **Bokeh** – Creación de dashboards, visualizaciones interactivas para aplicaciones web.

# Librerías de Python para DS

- **os** – operaciones con archivos del sistema.
- **Networkx** – Creación y manejo de información por medio de grafos/redes.
- **Re** – Expresiones regulares para encontrar ciertos patrones en textos.
- **Requests** – Envío de peticiones HTTP (conexión con APIs).
- **Beautiful Soup** – Scraping de páginas web.

# Setup



**ANACONDA<sup>®</sup>**

# Cursos recomendados

- **UDEMY**

- Python A-Z™: Python For Data Science With Real Exercises!
- The Complete SQL Bootcamp

- **COURSERA**

- Applied Data Science with Python Specialization
- Machine Learning (Andrew NG)
- Deep Learning Specialization

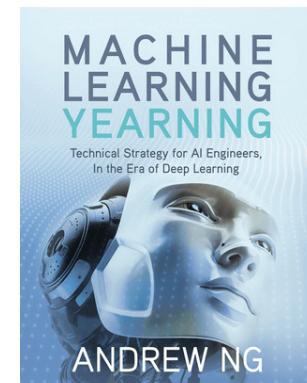
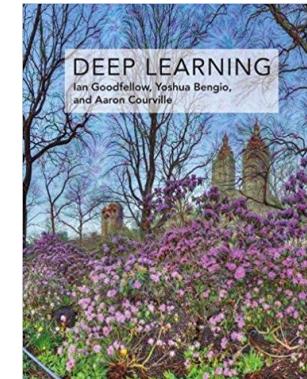
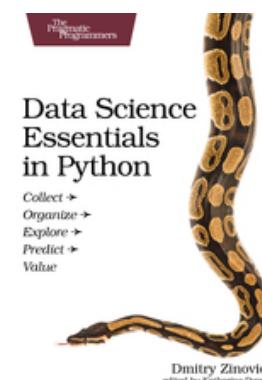
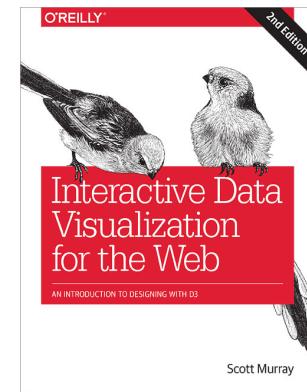
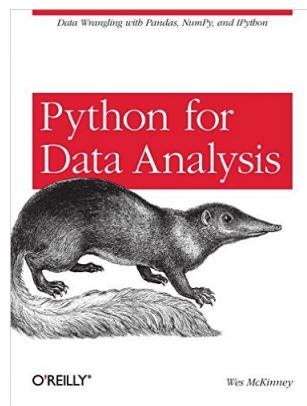


- **REGEX**

- <https://developers.google.com/edu/python/regular-expressions>

# Libros recomendados

- Python for Data Analysis
- Introduction to Machine Learning with Python: A Guide for Data Scientists
- Interactive Data Visualization for the Web, 2nd Ed.
- Data Science Essentials in Python
- Machine Learning Yearning
- Deep Learning



# Libros gratuitos de O'Reilly

- <http://www.oreilly.com/data/free>