Información Relevante y Confiable, por favor

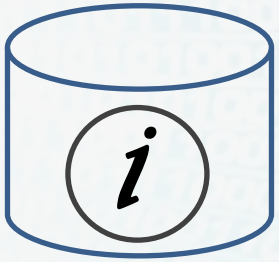
Marcelo Ferreyra

dataxplore

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

La Información

El Mundo
genera Datos

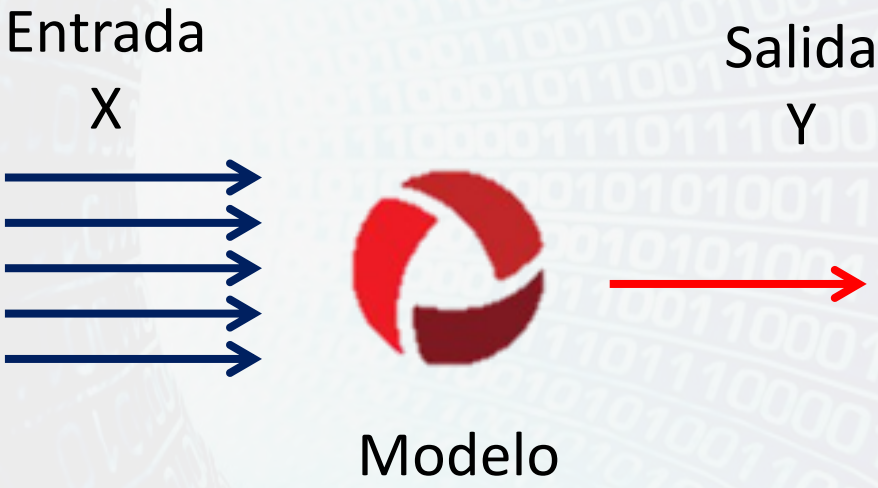


Los Datos
contienen
Información

Los Modelos
codifican esa
Información

La Información

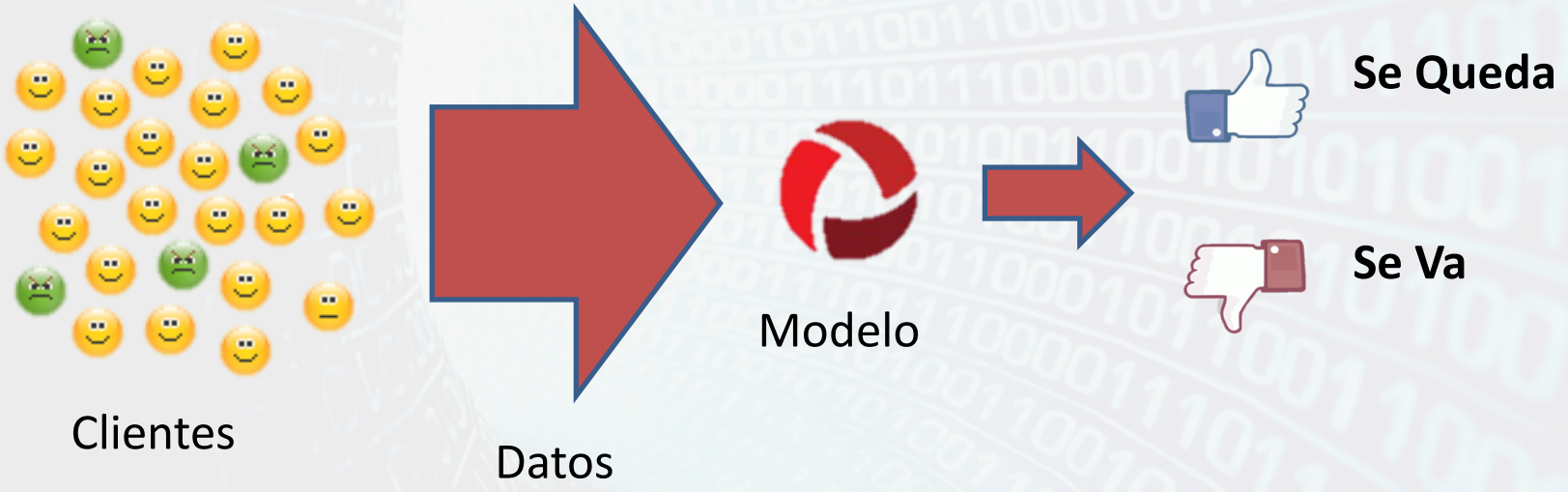
El modelo es un canal de información



El Modelo codifica la información proveniente de las variables de Entrada y transmite un mensaje hacia la Salida

La Información

Ejemplo: un modelo de Attrition



La Información

Los datos contienen Información

El modelo canaliza la Información

La información se utiliza para tomar una decisión

¿Qué es la Información?

La Información

Teoría de la Información de Claude Shannon

$$H = -\sum_{i}^{n} p_i * \log_2(p_i)$$

Entropía mide la Cantidad de Información. Se mide en BITS

Probabilidad de que arribe el mensaje ***i***

La Información

Ejemplo: una moneda

$$H = -\sum_{i}^{n} p_i * \log_2(p_i)$$

Lanzamiento de una moneda			
Mensaje	p	$-\log_2(p)$	$-p*\log_2(p)$
Cara	0.5	1	0.5
Cruz	0.5	1	0.5
		H	1

La Información

Ejemplo: una moneda cargada

$$H = -\sum_{i}^{n} p_i * \log_2(p_i)$$

Lanzamiento de una moneda			
Mensaje	p	$-\log_2(p)$	$-p*\log_2(p)$
Cara	0.3	1.737	0.521
Cruz	0.7	0.515	0.36
		H	0.881

Si las probabilidades no son iguales la entropía es menor

La Información

Cada variable contiene una determinada cantidad de información

Pero es mucho más interesante y útil conocer qué cantidad de información lleva una variable sobre otra

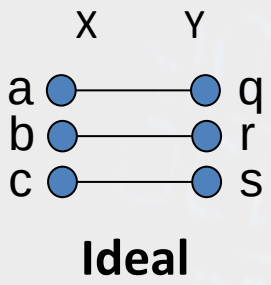
Supongamos una variable X que tiene tres valores distintos

a b c

Y queremos saber cómo está relacionada con otra variable Y que también tiene 3 valores distintos

q r s

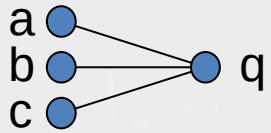
La Información

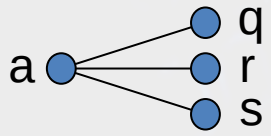


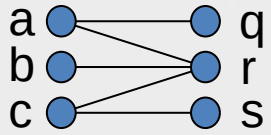
Ideal

El caso más simple es cuando hay una relación biunívoca entre las señales de X e Y

En este caso la información se transmite sin interferencias

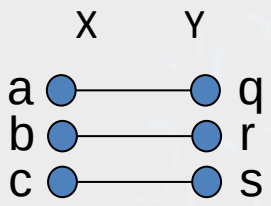


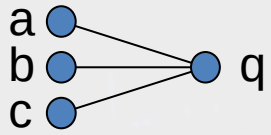




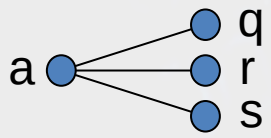
X \ Y	q	r	s
a	1.0		
b		1.0	
c			1.0

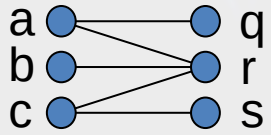
La Información





Equivocación

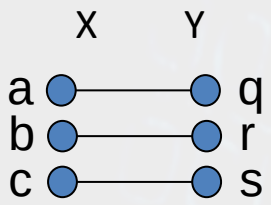


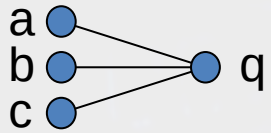


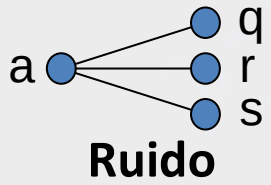
El segundo caso es cuando varias señales distintas de entrada apuntan a una sola señal de salida.
Muchas *voces* distintas están diciendo lo mismo.

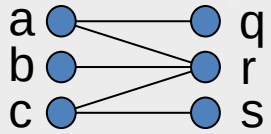
		Y: q	Y: r	Y: s
X	a	1.0		
X	b	1.0		
X	c	1.0		

La Información





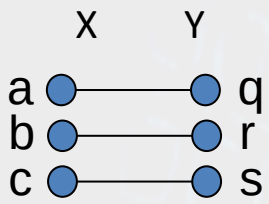


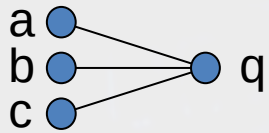


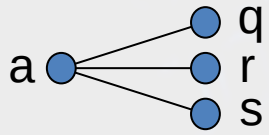
Cuando la relación contiene Ruido, una misma señal de entrada apunta a distintas señales de salida.
La señal de salida es incierta para una determinada señal de entrada

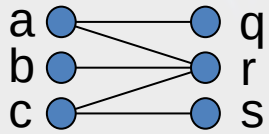
		Y		
		q	r	s
X	a	0.3	0.3	0.3
	b			
	c			

La Información









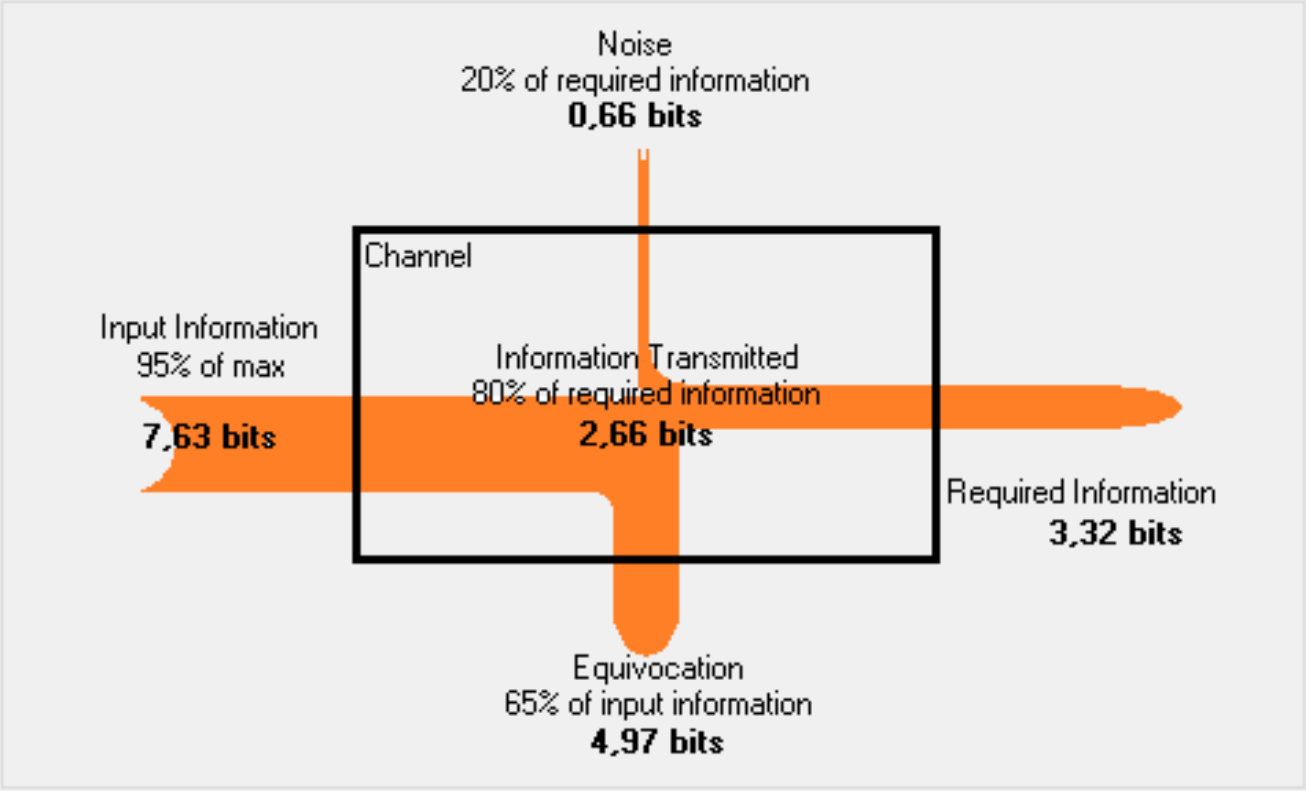
Ruido y Equivocación

Con datos reales lo normal es que exista una mezcla de información, ruido y equivocación.

		Y		
		q	r	s
X	a	0.5	0.5	
	b		1.0	
	c		0.5	0.5

La Información

Ejemplo: medición de Información en una base de datos



Medido con Powerhouse Analytics

La Información

Ventajas de medir la información

- Es posible conocer si los datos contienen información aún antes de modelar
- El ruido tiene una definición precisa
- Se obtiene una referencia con la que comparar el modelo
- Se puede utilizar para seleccionar variables

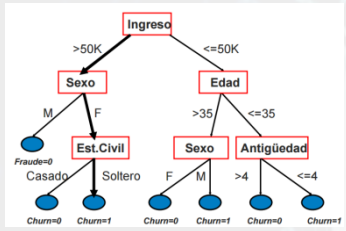
La Información

Selección de variables

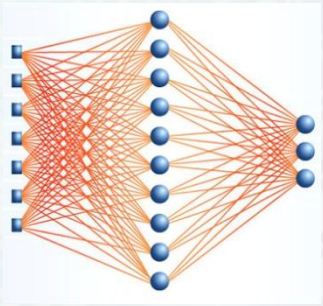
1. Seleccionar la variable que mayor información contenga acerca de la variable a predecir
2. Seleccionar la siguiente variable con mayor información *adicional* acerca de la variable a predecir
3. Continuar con el paso 2 hasta que la cantidad de información que aporte la variable no justifique la pérdida de representatividad

La Información

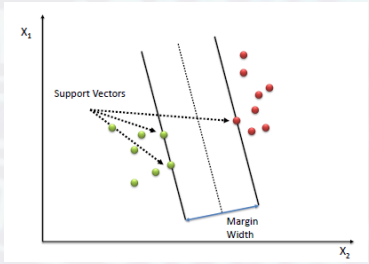
Las herramientas de DM son muy útiles para obtener información



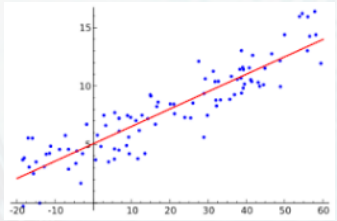
Árboles de Decisión



NN



SVM



Regresión Lineal

Pero no alcanzan para conocer su relevancia

Información Relevante

Trabajo multidisciplinario

- IT o quienes administran los Datos
- Data Miners o Data Scientist
- Quienes conocen el Negocio
- Científicos Sociales

Información Relevante

¿Por qué científicos sociales?

Antropólogos y Sociólogos cuentan con herramientas de investigación y análisis necesarias para comprender los distintos códigos culturales que forman parte de la sociedad.

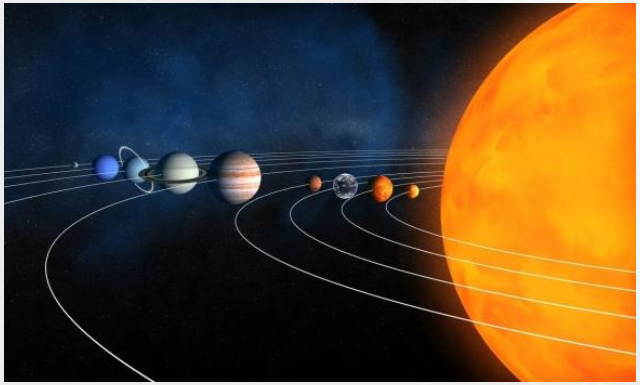
A través de la *desnaturalización*, acostumbran a explorar los supuestos para dejarlos al descubierto

La comprensión del *otro* para poder ver con más claridad los distintos segmentos socioeconómicos tanto desde adentro como desde afuera

Información Relevante

Modelos empíricos versus teóricos

Johannes Kepler versus Isaac Newton



Kepler desarrolló un modelo empírico que permite calcular los movimientos planetarios

Newton propuso un modelo teórico con el que se puede entender y predecir

Información Relevante

Utilizar el *Método Científico*

La ciencia es más una determinada manera de pensar, que un cuerpo de conocimientos

Carl Sagan
El Cerebro de Broca

Información Relevante

Ejemplo de un marco teórico

Mercados Financieros

Hipótesis del Mercado Eficiente

Los precios reflejan toda la información pública y siguen un *camino al azar*

Los inversores son *racionales*

Teoría del Caos

Los precios muestran comportamientos de sistemas dinámicos no lineales

dataxplore

Información Confiable

Desarrollo versus Producción

- Cantidad de datos necesarios
- Uso de la muestra de validación
- Pruebas estadísticas
- Correlaciones

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

dataxplore

Información Confiable

Cantidad de datos necesarios

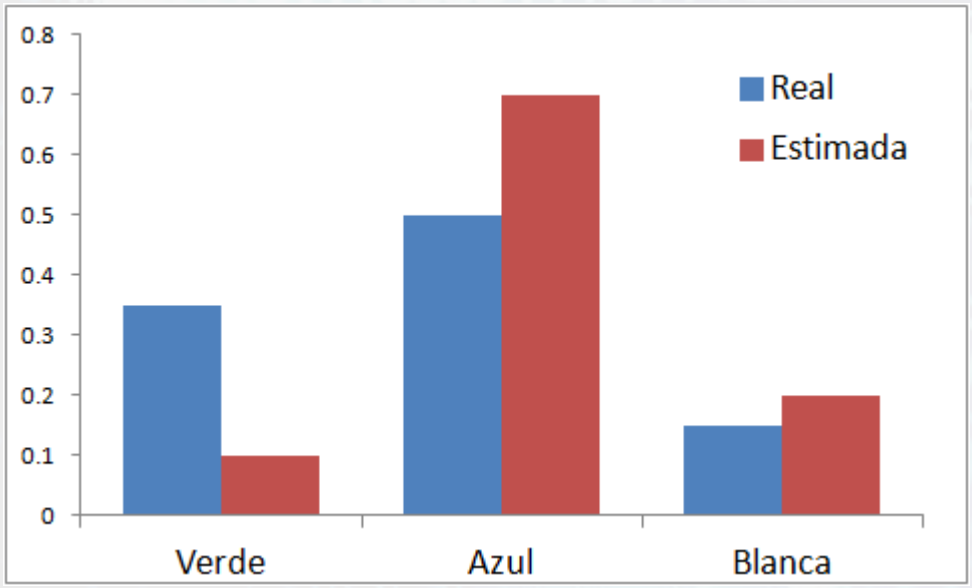
Problema: una urna contiene bolillas de colores.
¿Cuántas bolillas al azar se deben sacar para estimar la distribución?

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Información Confiable

Cantidad de datos necesarios

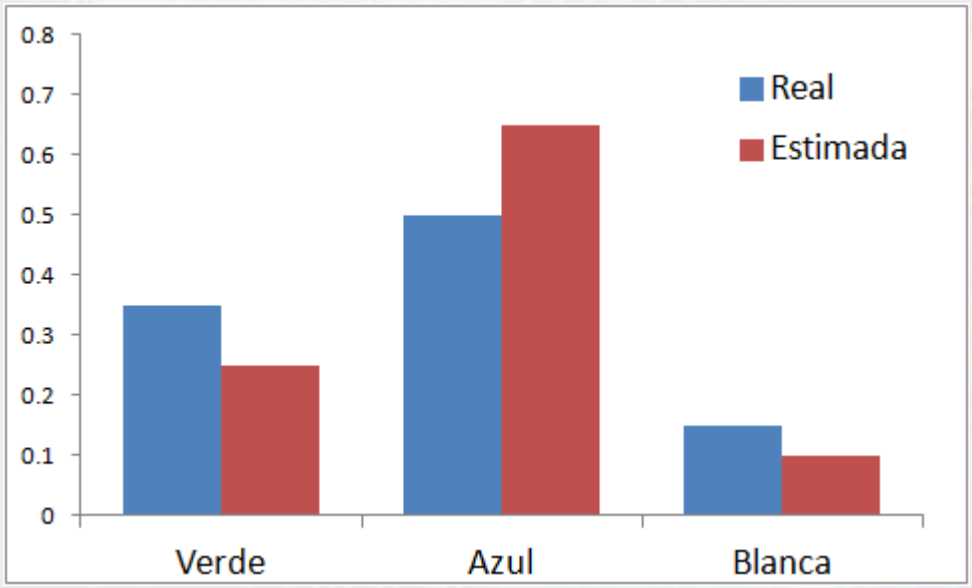
Problema: una urna contiene bolillas de colores.
¿Cuántas bolillas al azar se deben sacar para estimar la distribución?



Información Confiable

Cantidad de datos necesarios

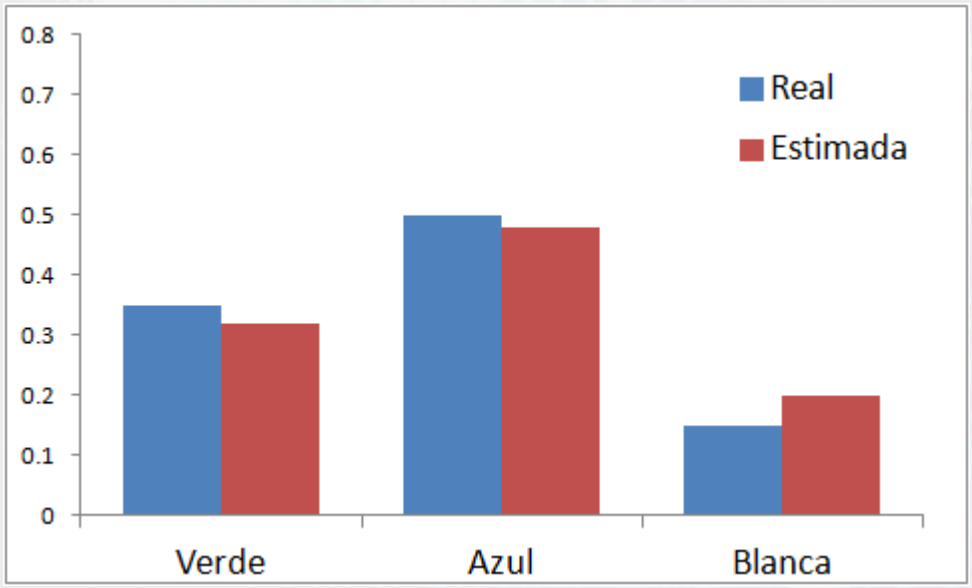
Problema: una urna contiene bolillas de colores.
¿Cuántas bolillas al azar se deben sacar para estimar la distribución?



Información Confiable

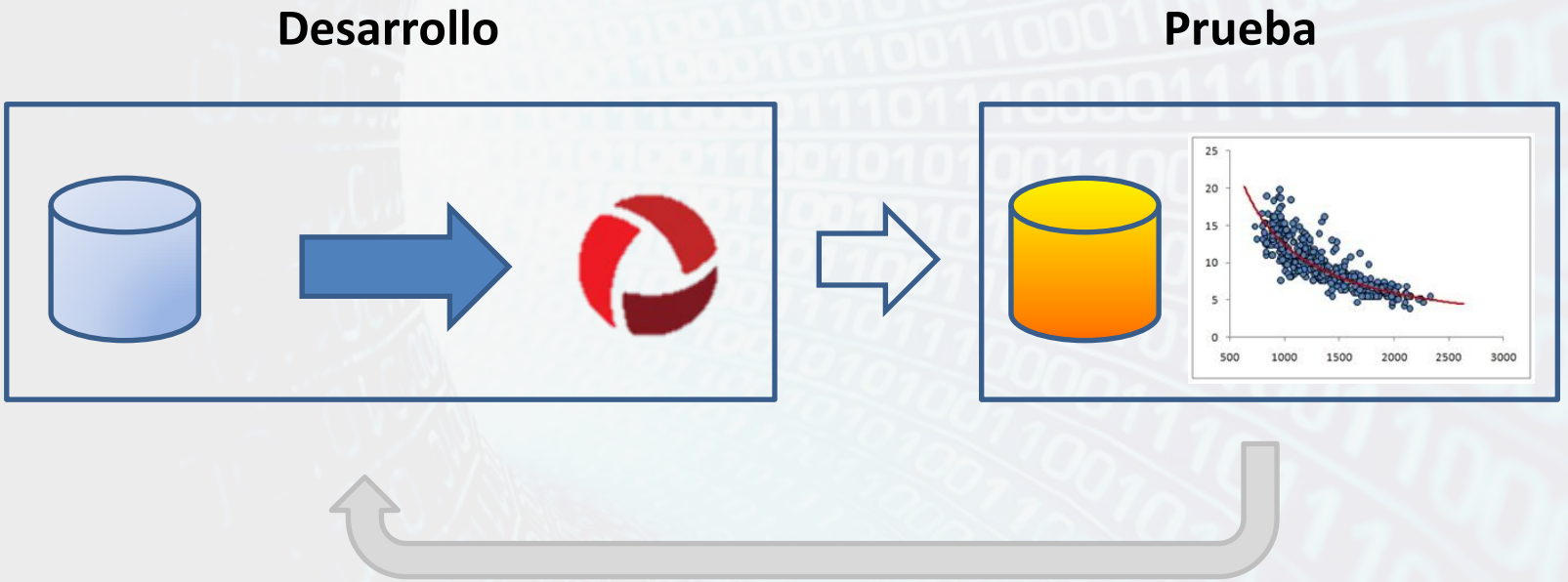
Cantidad de datos necesarios

Problema: una urna contiene bolillas de colores.
¿Cuántas bolillas al azar se deben sacar para estimar la distribución?



Información Confiable

Uso de la muestra de prueba



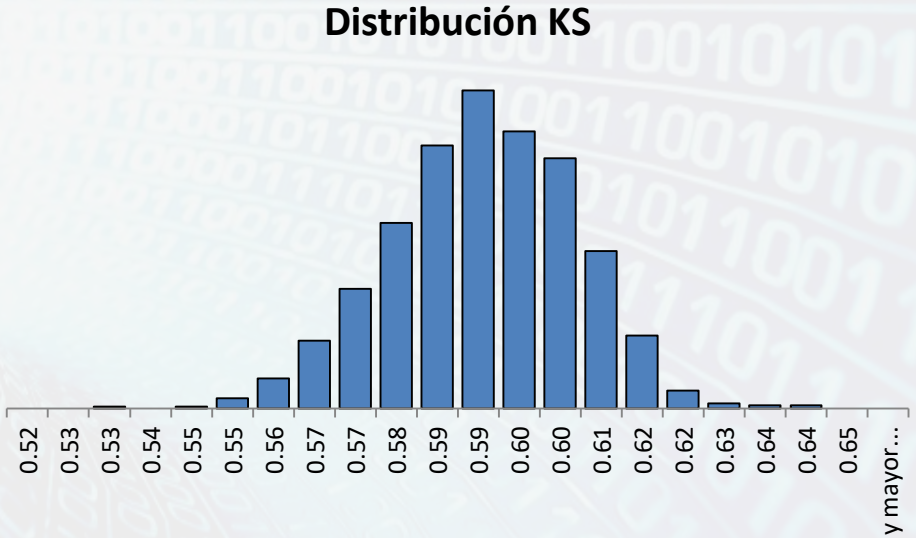
Información Confiable

Pruebas estadísticas

Es necesario realizar las pruebas estadísticas apropiadas para validar los modelos

KS validado con Bootstrap

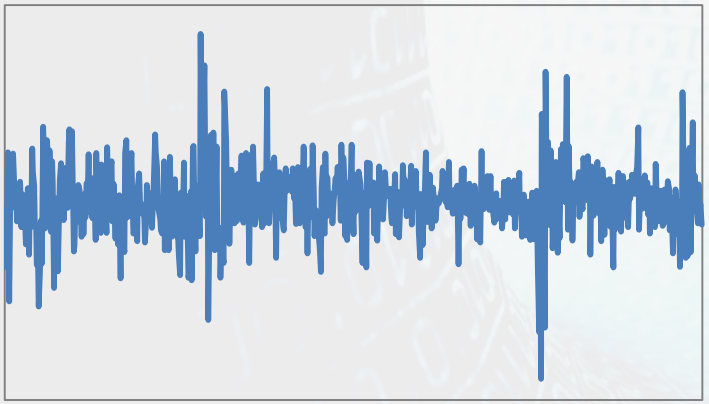
KS = 58.9
55.9 < KS < 61.5 (95%)



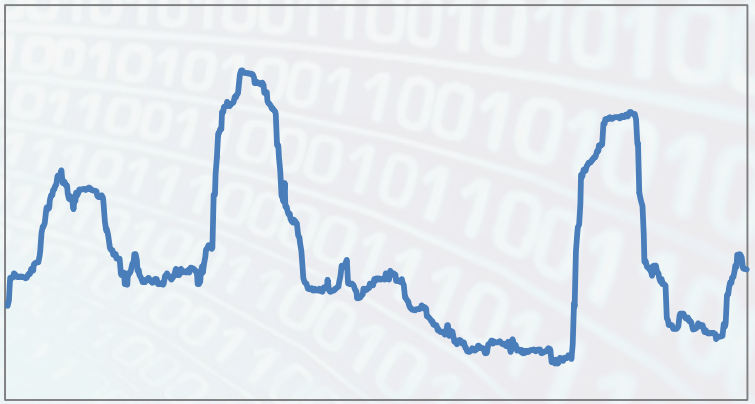
Información Confiable

Estabilidad de la distribución de los datos

El mercado financiero no tiene una distribución estable a lo largo del tiempo



Cambios de Precios

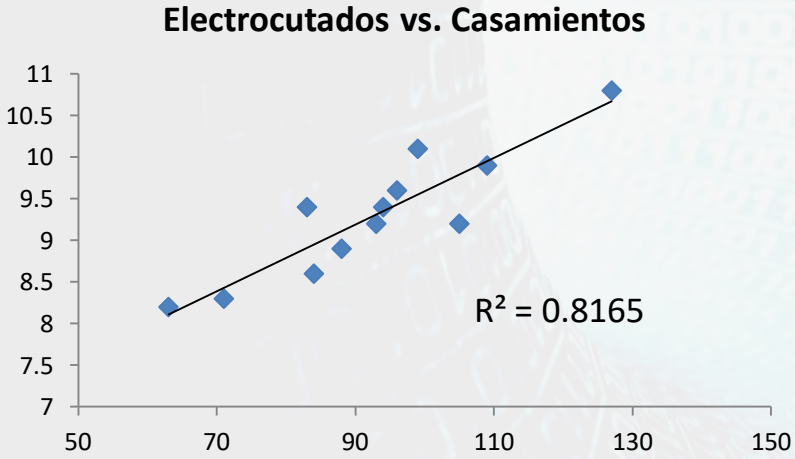


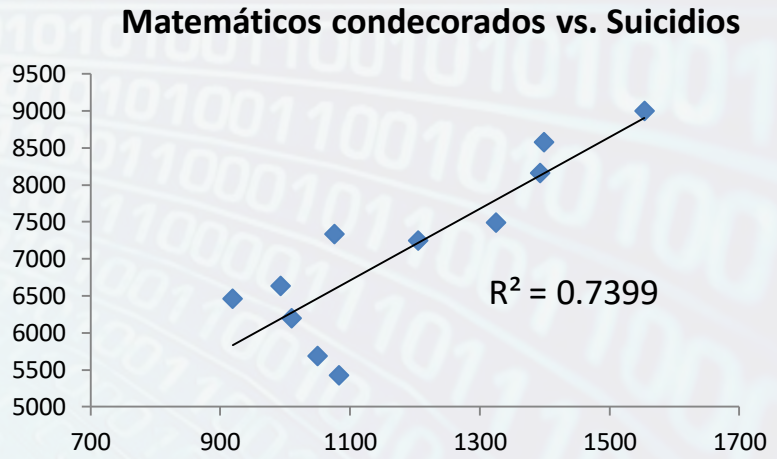
Varianza

Información Confiable

Correlaciones: Coincidencia de cosas que ocurren juntas

Correlaciones espureas





Información Confiable

Correlaciones

Teoría de Ramsey (1903-1930): El desorden total es imposible

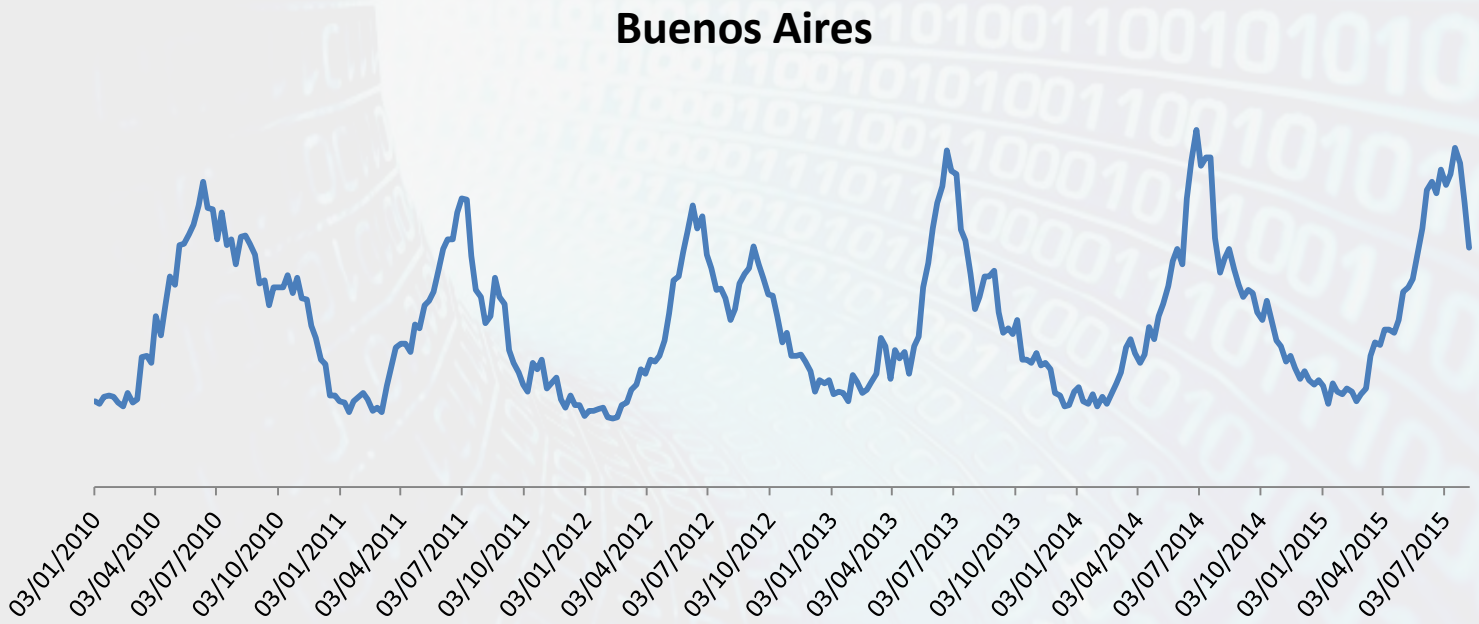
La probabilidad de encontrar correlaciones espureas en una base de datos aumenta con el tamaño de la misma

"The Deluge of Spurious Correlations in Big Data"
C. Calude and G. Longo

Información Confiable

Google Flu Trends

Google descubrió que las búsquedas de enfermedades en Internet se adelantaban a las epidemias de gripe. En 2008 desarrollaron un modelo y lo actualizaron en 2009



Información Confiable

Google Flu Trends

En el invierno 2011-2012 GFT sobrestimó por más del 50% el número de casos de gripe reportados por el Centro de Control y Prevención de Enfermedades de los Estados Unidos

Sucedió algo similar durante el invierno 2012-2013

En Agosto de 2015, Google dejó de publicar predicciones de GFT

dataxplore

Resumen

Medir la Información en los Datos

Proyectos multidisciplinarios incorporando Científicos Sociales

Buscar la Teoría que respalde la Información encontrada

Reforzar las buenas prácticas

UNIVERSIDAD AUSTRAL
Facultad de Ingeniería

Referencias

Data Mining basado en Teoría de la Información:

http://web.austral.edu.ar/images/contenido/facultad-ingenieria/2-Data_Mining_basado_Teoria_Informacion_Marcelo_Ferreyra.pdf

http://powerhousedm.blogspot.com.ar/

http://www.dataxplore.com.ar/tecnologia.php#Powerhouse

Google Flu Trends:

https://www.google.org/flutrends/about/