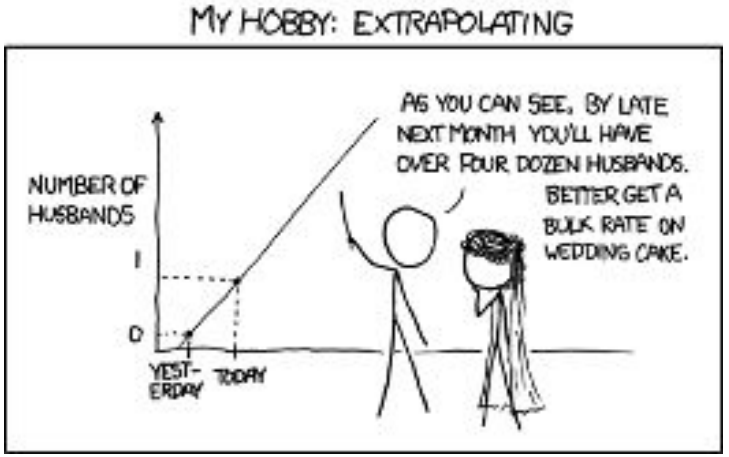
Límites Estadísticos de la Minería de Datos

Andrés Abeliuk

Significado de las respuestas

- Un riesgo de la minería de datos es que "descubrirás" patrones que no tienen sentido.
- Los estadísticos lo llaman el **principio de Bonferroni:**
  Si buscas patrones interesantes en más lugares de los que tú cantidad de datos admite, seguramente encontrarás basura.
- El principio de Bonferroni nos ayuda a distinguir estas ocurrencias aleatorias y evitar tratarlas como eventos reales.



Principio de Bonferroni

- Las tecnologías de minería de datos y los datos masivos hacen que sea increíblemente fácil hacer preguntas y probar hipótesis.
- es decir, puedes ejecutar de forma eficaz cientos de experimentos en unas pocas horas. Obtendrás muchos "positivos" por casualidad.
- Es muy importante saber cuál es la tasa de falsos positivos y si un resultado que ves es realmente "inusual" en relación con el azar.
- **Buena práctica**: informa todo lo que intentas en tus experimentos, no solo los éxitos.

Principio de Bonferroni

- El principio de Bonferroni es una presentación informal de un teorema estadístico:
  - Si su método para encontrar elementos interesantes **devuelve significativamente más elementos** de los que esperaría en la población real,
  - puede asumir que **la mayoría de los elementos** que encuentra con él son **falsos**.
- Para esto necesitamos primero calcular el número esperado de ocurrencias de los eventos que se están buscando suponiendo que los datos son completamente aleatorios.
- La aplicación del principio de Bonferroni a un algoritmo para identificar o clasificar datos te **da un límite superior en la precisión de tus métodos.**

Rajaraman A, Leskovec J, Ullman J.*Mining of Massive Datasets* Version 1.3

Ejemplo de Principio de Bonferroni: Paradoja de Rhine

- Total Information Awareness (TIA) (**búsqueda de actividad sospechosa**), US IAO 2003
- **Vigilancia predictiva**: el uso de técnicas matemáticas, predictivas y analíticas para identificar una posible actividad delictiva.
- En 2002, la administración Bush presentó un plan para extraer todos los datos que pueda encontrar, incluidos recibos de tarjetas de crédito, registros de hoteles y datos de viajes para rastrear la actividad terrorista.
- Este proyecto fue (oficialmente) destrozado por el Congreso
- Una gran objeción a TIA fue que estaba **buscando tantas conexiones vagas** que seguramente **encontraría cosas que eran falsas** y, por lo tanto, **violaban la privacidad de personas inocentes.**

Un ejemplo simplificado para ilustrar problemas potenciales con TIA

- Definición del problema:
  - Supongamos que los "terroristas" se reúnen periódicamente en un hotel para planear sus actividades. Queremos detectarlos basándonos en los siguientes supuestos:
- Supuestos:
  - Hay mil millones de personas (10^9) que podrían ser terroristas.
  - Todo el mundo va a un hotel un día de cada 100.
  - Un hotel tiene capacidad para 100 personas.
  - Examinaremos los registros de hoteles durante 1000 días.
- Para encontrar “terroristas”, buscaremos personas que, en dos días diferentes, estuvieron ambos en el mismo hotel.

Si todos se comportan al azar (es decir, no hay terroristas), ¿la minería de datos detectará algo sospechoso?

Cálculo de Sospechosos

- 10^9 personas siendo rastreadas
- 1000 días
- Cada persona permanece en un hotel el 1% del tiempo (10 días de cada 1000).
- Los hoteles tienen capacidad para 100 personas
- 10^5 hoteles para albergar al 1% de 10^9 personas

- Probabilidad de que una persona **p** y **q** estén en el mismo hotel un día **d**:

$$1/100 \times 1/100 \times 10^{-5} = 10^{-9}$$

- Probabilidad de que **p** y **q** estén en el mismo hotel los días **d1** y **d2** dados:

$$10^{-9} \times 10^{-9} = 10^{-18} \quad \textbf{(A)}$$

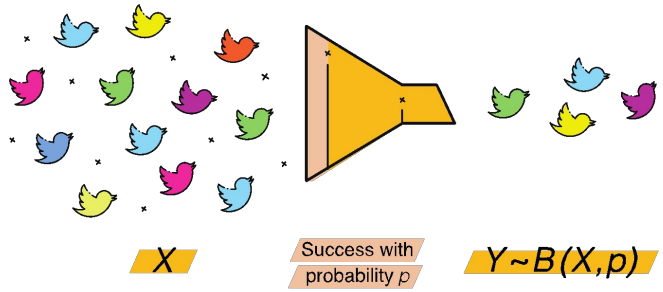
- Pares de días: $1000C2$ que es approx. 5×10^5 **(B)**
- Probabilidad de que **p** y **q** estén en el mismo hotel dos días:
  - (número de pares de días) x (prob p y q en el mismo hotel durante 2 días)
  - **(B)** X **(A)** = $(5 \times 10^5) \times 10^{-18} = 5 \times 10^{-13}$ **(C)**
- Pares de personas: 10^9C2 approx. $5x10^{17}$ **(D)**
- Número esperado de parejas de personas "sospechosas":
  - **(C)** X **(D)** = $5 \times 10^{-13} \times 5x10^{17}$ = **250.000**

Límites teóricos de la precisión predictiva

Pronosticar fenómenos complejos

- Pronosticar desde **epidemias** hasta **opiniones públicas**, **mercado de valores** y **ataques cibernéticos**, es fundamental para muchas aplicaciones.
- La predicción es además el marco estándar para evaluar modelos de sistemas complejos aprendidos de los datos.
- El pronóstico de series de tiempo, que se usa ampliamente para modelar fenómenos dinámicos, representa un proceso como una secuencia de observaciones (recuentos de eventos discretos o continuos) a intervalos de tiempo regulares.
- Se han desarrollado modelos de predicción basados en **procesos puntuales estocásticos y auto-excitantes**, **modelos de Markov** y **modelos autorregresivos**, para predecir
  - delitos, malestar social, terrorismo, epidemias, movilidad humana, correspondencia personal, actividad en línea, sistemas ecológicos, ...

Los sistemas complejos rara vez se observan completamente



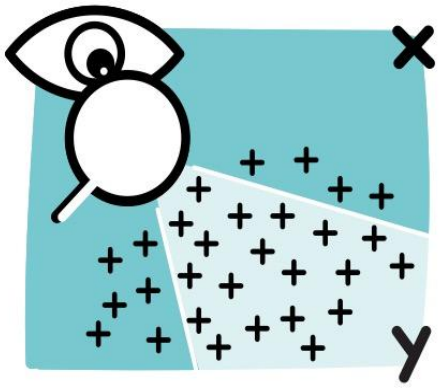
Un desafío fundamental para los esfuerzos de modelamiento y predicción es el hecho de que los sistemas complejos rara vez se observan completamente.

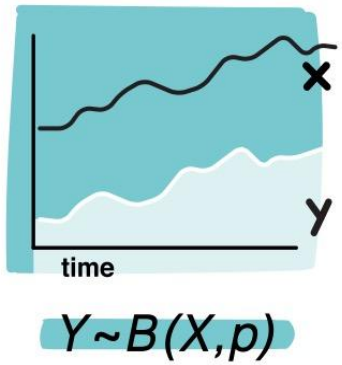
- ej. Twitter muestra del 10%; Encuestas temporales; falta de capacidad de prueba, ...

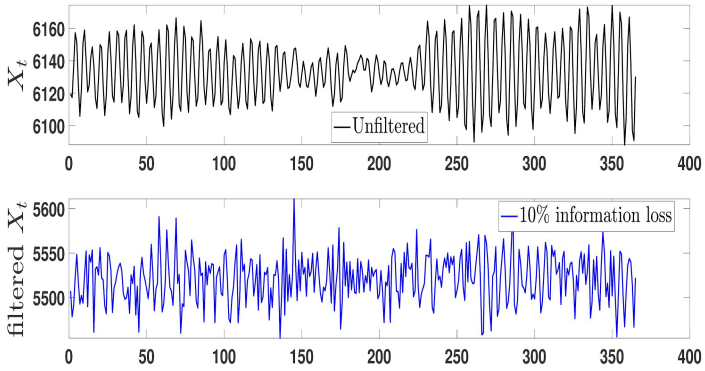
¿El muestreo temporal introduce sesgos en el fenómeno dinámico?

¿Se pueden aprender modelos precisos a partir de observaciones incompletas?

Modelo







- Considere una serie de tiempo con conteos diarios de algún evento
- Representamos el proceso como una serie temporal de conteos, $X=[X_1,\ldots,X_t]$
- Es posible que los observadores de este proceso no vean todos los eventos. Nos referimos a la serie temporal de eventos observados, $Y=[Y_1,\ldots,Y_t]$
- La probabilidad de que se observe un evento es p. La distribución binomial B(X,p) se utiliza para modelar la señal observada Y, i.e.,

$$Y_t \sim B(X_t, p) \quad \forall t$$

Abeliuk, A., Huang, Z., Ferrara, E. *et al.* Predictability limit of partially observed systems. *Sci Rep*, 2020

Cuantificación de la predictibilidad

Cuantificación sin modelos de la predictibilidad de series de tiempo

- En el nivel más simple, la **autocorrelación** captura qué tan bien se correlaciona una serie de tiempo con sus propias versiones desplazadas en el tiempo.
- En ecología y física, la **entropía de permutación** se utiliza para medir la predictibilidad. La entropía de permutación (PE) captura la complejidad de una serie de tiempo a través de estadísticas de sus subsecuencias ordenadas..

Permutation Entropy (PE)

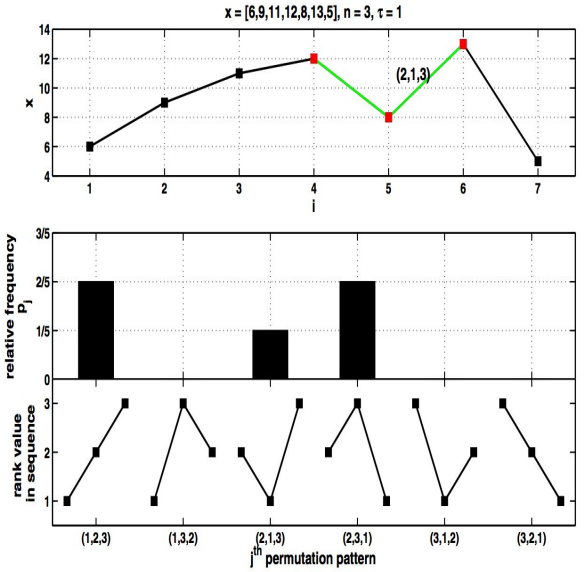


Image source: Müller et al. The European Physical Journal Special Topics, 2013

$$H^{\mathrm{P}}(d) = -\sum_{\pi \in \mathcal{S}_d} P(\pi) \log_2 P(\pi)$$

Cuantificación de la pérdida de predictibilidad

Lemma:

Assume X_t, Y_t are two time series related by the relation $Y_t \sim B([X_t], p)$. The covariance matrices $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ are related as

$$\mathrm{Cov}(Y_i, Y_j) \approx p^2 \mathrm{Cov}(X_i, X_j)$$

$$\mathrm{Var}[Y_t] \approx p(1-p)E[X_t] + p^2 \mathrm{Var}[X_t]$$

Key: An approximation to B(n, p) is given by the normal distribution $\mathcal{N}(np, np(1-p))$

Cuantificación de la pérdida de predictibilidad

Nuestra principal contribución teórica es una caracterización analítica del impacto del muestreo en la autocorrelación de una señal.

La autocorrelación se define como la correlación de Pearson entre los valores rezagados de la señal, es decir,

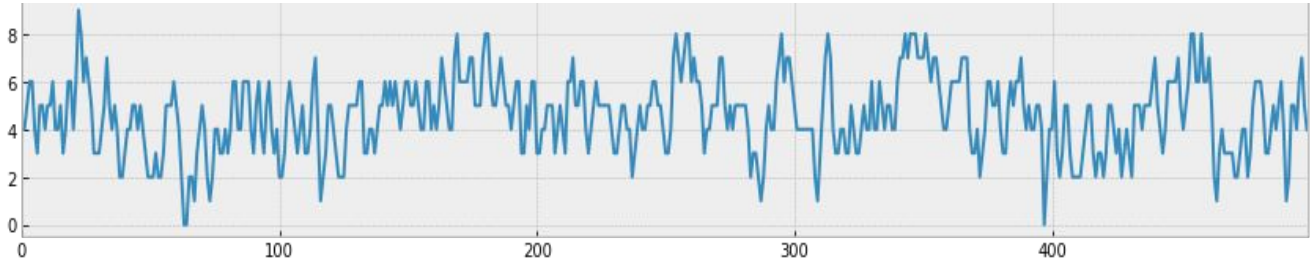
$$\rho_{Y_i,Y_j} = \frac{Cov(Y_i, Y_j)}{\sigma_{Y_i}\sigma_{Y_j}} \approx \frac{p^2 \mathrm{Cov}(X_i, X_j)}{p^2 \mathrm{Var}(X_t) + p(1-p)E[X_t]}$$

Corollary 1: **Decaimiento de la autocorrelación de la señal observada.**
La autocorrelación de la señal observada Y decae monótonamente a velocidades de muestreo más bajas.

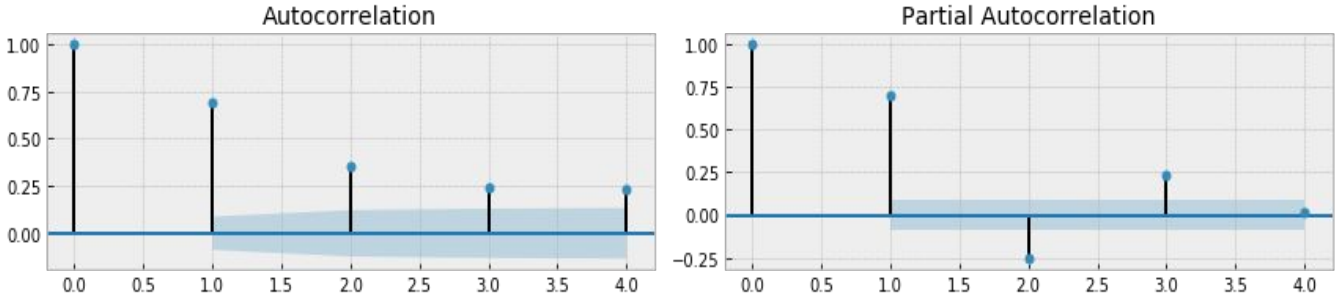
Corollary 2: **Decaimiento de la covarianza con la señal externa.**
La correlación entre las señales observadas y externas se degrada linealmente a velocidades de muestreo más bajas.

Ejemplo sintético

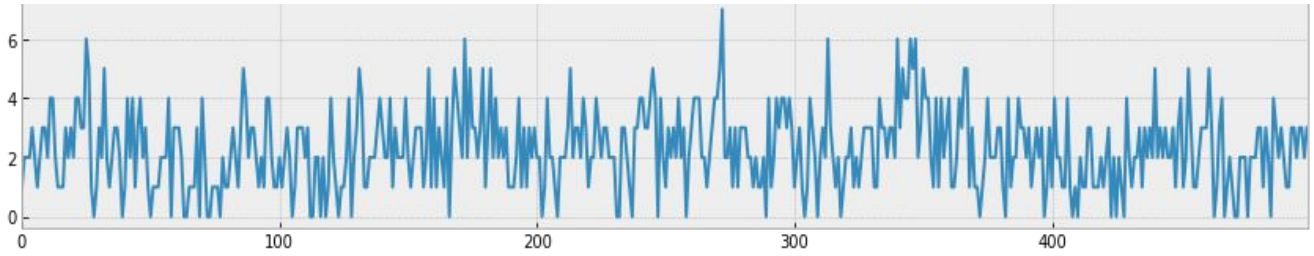
ARMA(3, 2) model simulated



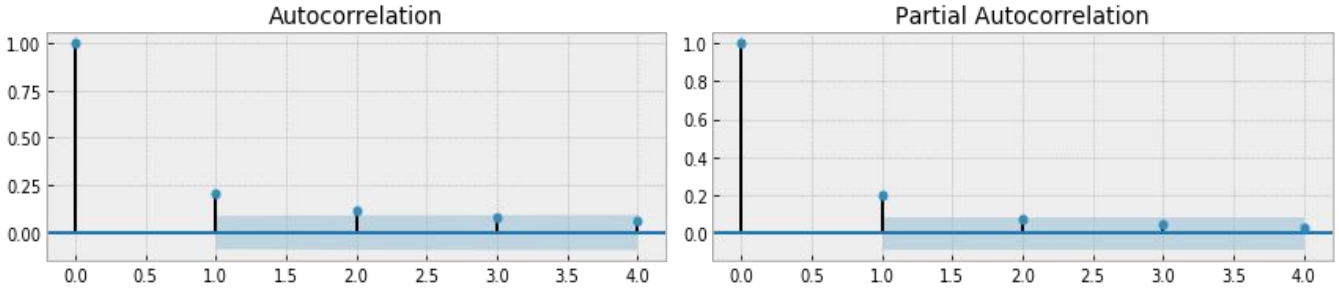
Autocorrelation between lagged counts



Filtered ARMA(3, 2) model with p=0.5

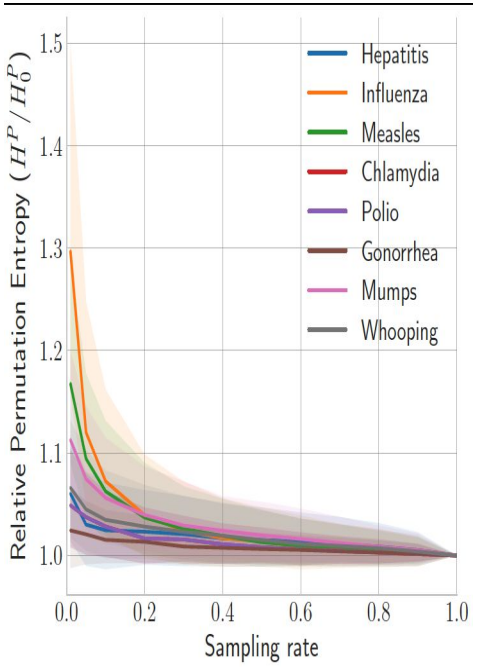


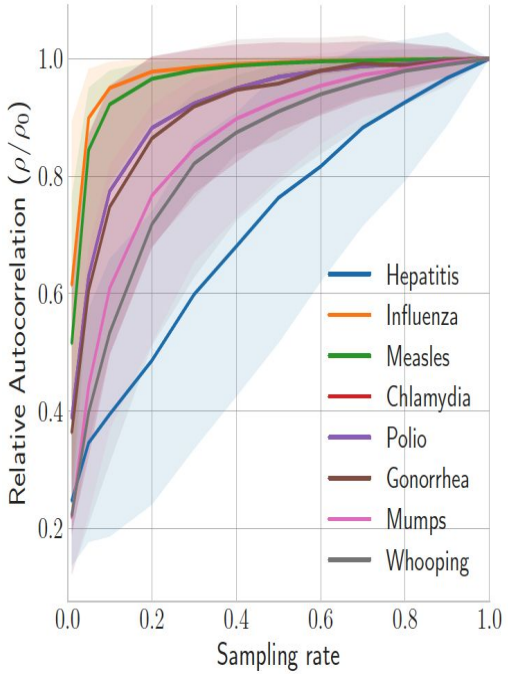
Autocorrelation is almost completely lost

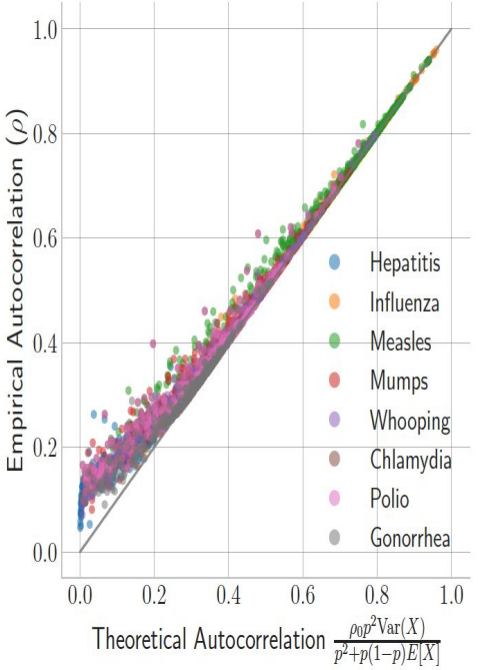


Datos del mundo real: Epidemias

Pérdida de predictibilidad en brotes de enfermedades debido al muestreo.

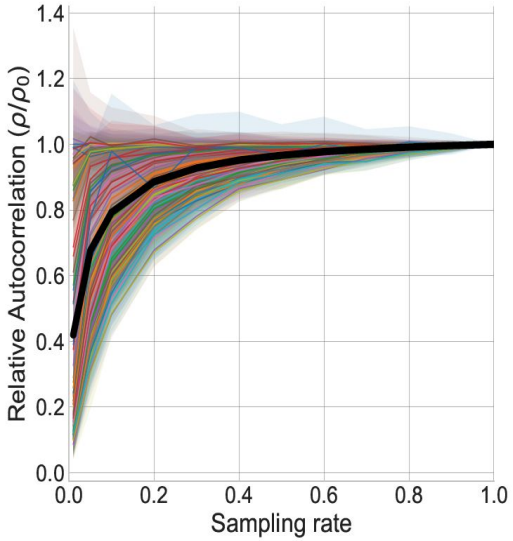


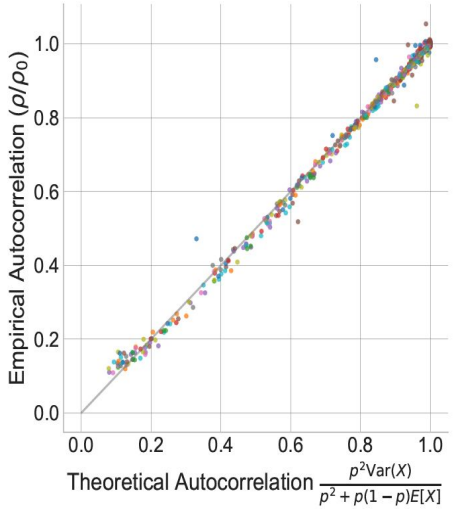


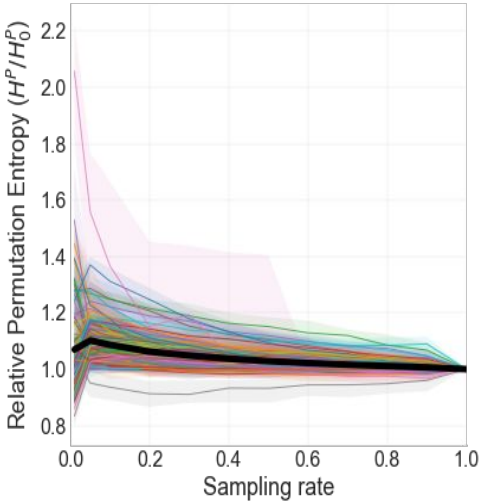


Estudiamos ocho enfermedades (clamidia, gonorrea, hepatitis A, influenza, sarampión, paperas, poliomielitis y tos ferina), que representan cada brote de enfermedad como una serie de tiempo del número semanal de infecciones notificadas en cada estado de EE. UU.

Datos del mundo real: Twitter

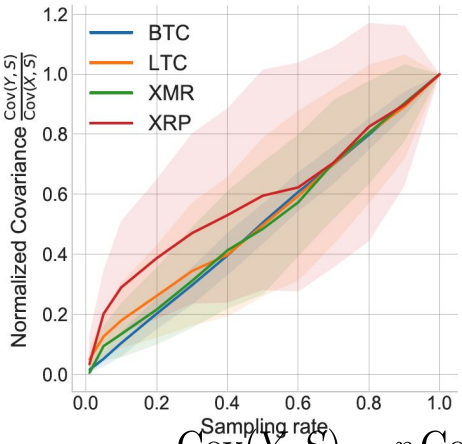




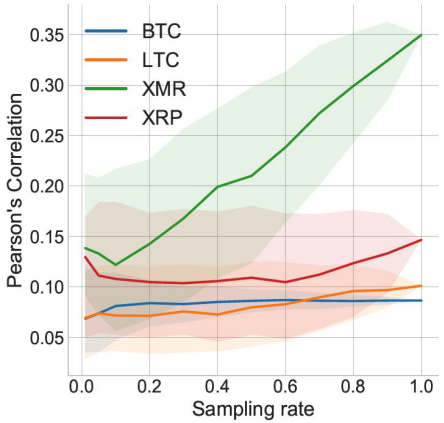


- Analizamos la predictibilidad de la popularidad de los hashtags, definida como el número diario de menciones del hashtag.
- Los efectos del muestreo a diferentes tasas sobre la autocorrelación y la precisión de la predicción.
  - **La fig. izquierda** muestra la pérdida promedio de autocorrelación relativa a la serie de tiempo original
  - **El diagrama central** muestra la precisión de las predicciones teóricas.
  - **El gráfico de la derecha** muestra un aumento de la impredictibilidad (entropía de permutación) debido al muestreo

Datos del mundo real: criptomonedas



$$\mathrm{Cov}(Y, S) = p\,\mathrm{Cov}(X, S)$$



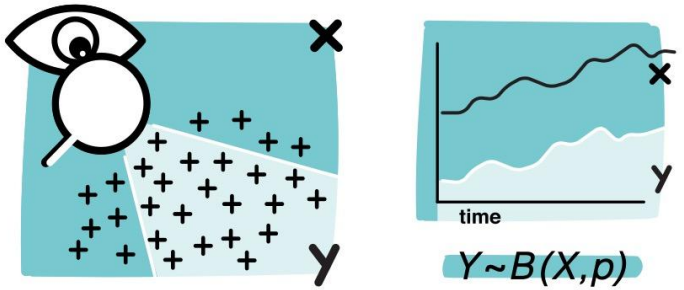
Cryptocurrency	Events	Repositories	Users	Events/Day (SD)	Events/Repo (SD)
Bitcoin (BTC)	40,038	1,962	5,324	460 (116)	20.4 (138.4)
Ripple (XRP)	2,963	7	86	35.3 (27.7)	423.2 (1,115)
Litecoin (LTC)	1,222	137	302	14 (11)	8.9 (19.2)
Monero (XMR)	370	15	54	5.7 (6.3)	24.7 (55.8)

Table 1: Github data: descriptive statistics.

Presentamos hallazgos con respecto a la pérdida de correlación entre una serie de tiempo muestreada y una señal externa.

- Pérdida de correlación entre la popularidad del repositorio de criptomonedas y sus precios para diferentes tasas de muestreo.
- **Fig. izquierda:** muestra una disminución en la covarianza relativa para tasas de muestreo más bajas, corroborando nuestros resultados teóricos.
- **Fig. derecha:** Vemos que una disminución de la covarianza tiende a inducir una pérdida de correlación, especialmente para aquellas monedas con baja variación en relación con su media.

Conclusiones



- Propusimos un marco teórico para analizar los efectos de tener solo información parcial sobre un fenómeno dinámico.
- Nuestro marco nos permite medir el impacto del muestreo temporal en la predictibilidad de un proceso.
- Mostramos que **el muestreo destruye información útil** sobre el proceso dinámico, como lo demuestra una **disminución en su autocorrelación y un aumento en la entropía de permutación**, en función de la frecuencia de muestreo p.

¿Utilizas datos parcialmente observables en tus estudios?