Big Data y Aprendizaje automático en Economía y Ciencias Sociales

Natalia da Silva

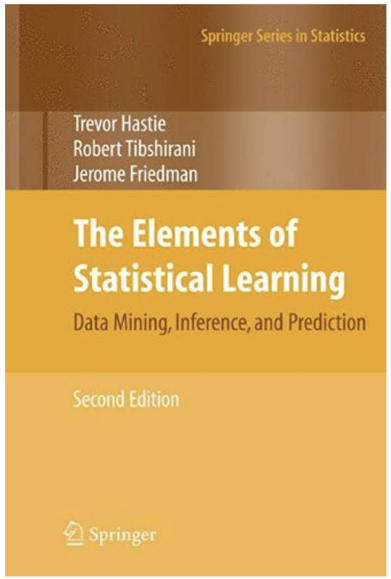
Instituto de Estadística-FCEA-UdelaR
XXXIV Jornadas Anuales de Economía

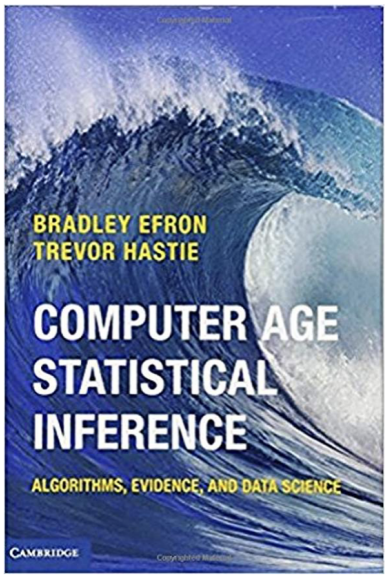
natalia@iesta.edu.uy - natydasilva.com - @pacocuak

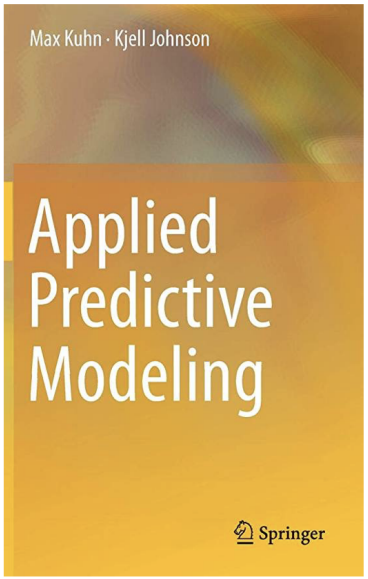
21 de Agosto

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Algo de material







Algunos términos

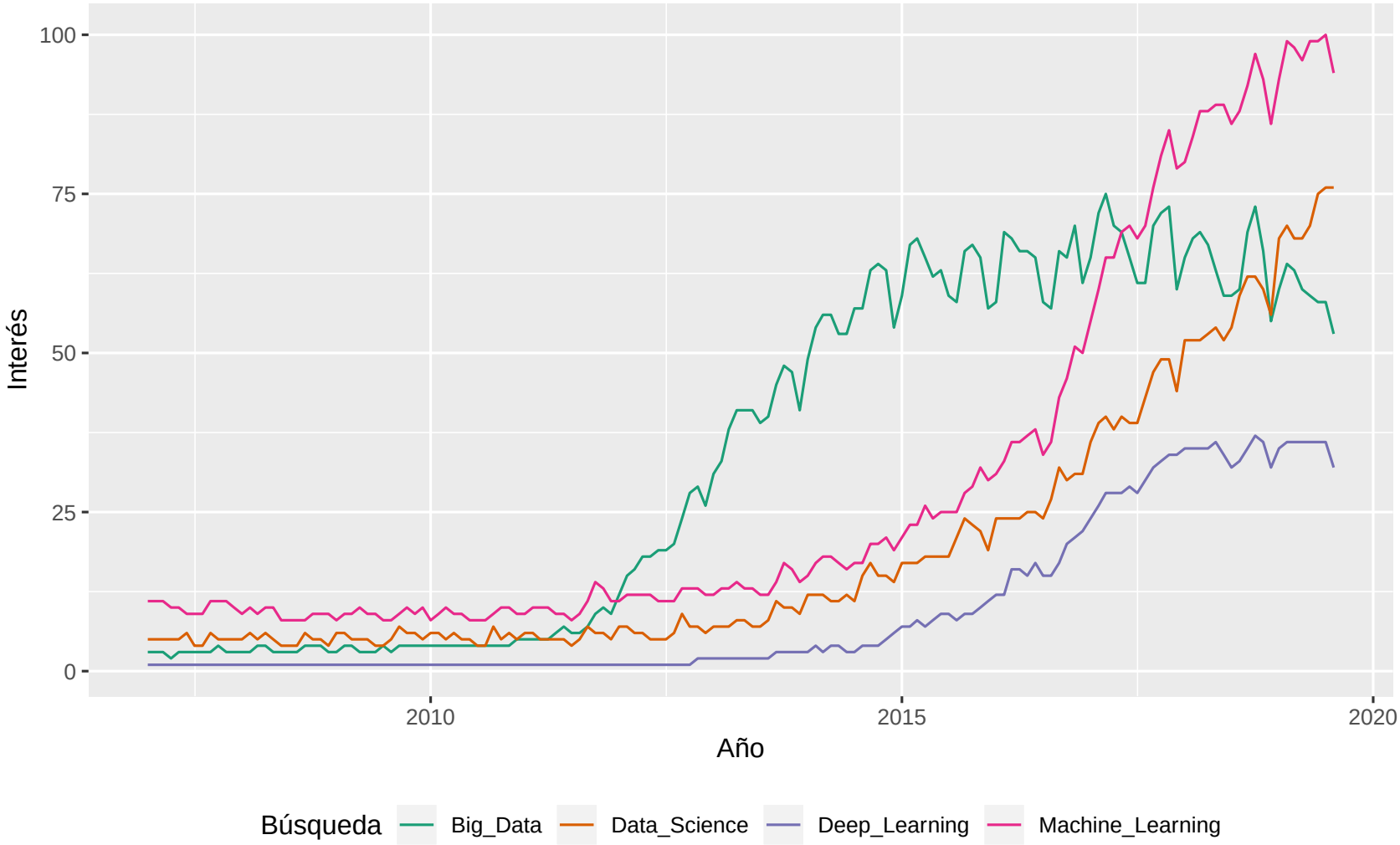
- Big data: datos complejos por Volumen, Variedad o Velocidad

- Aprendizaje automático: métodos y algoritmos para detectar patrones predecir nuevos datos.

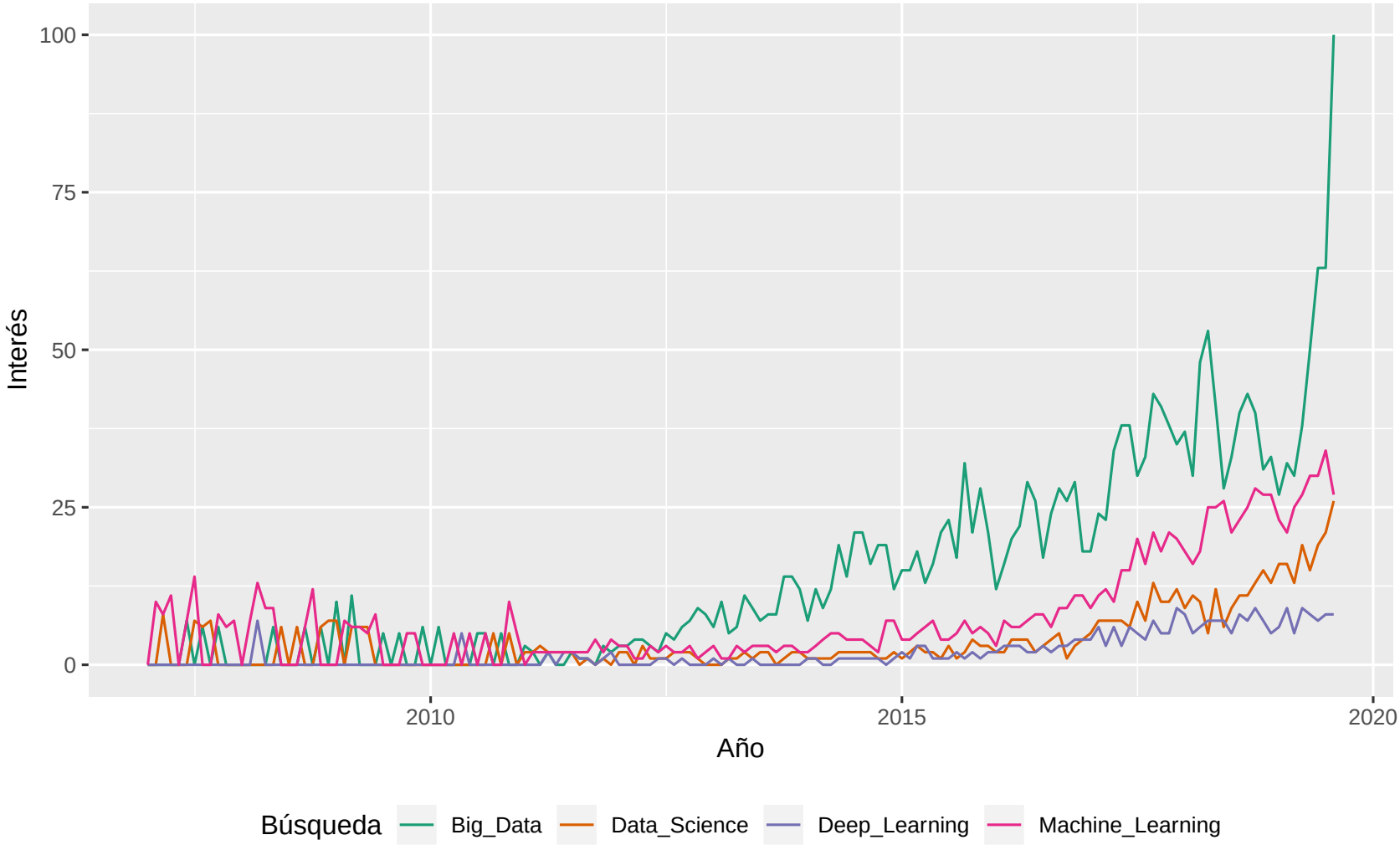
Más Términos

- **Big data**
- **Aprendizaje automático**
- **Ciencia de datos**
- **Reconocimiento de patrones**
- **Minería de datos**
- **Aprendizaje profundo**
- **Aprendizaje estadístico**
- **Inteligencia Artificial**
- .....

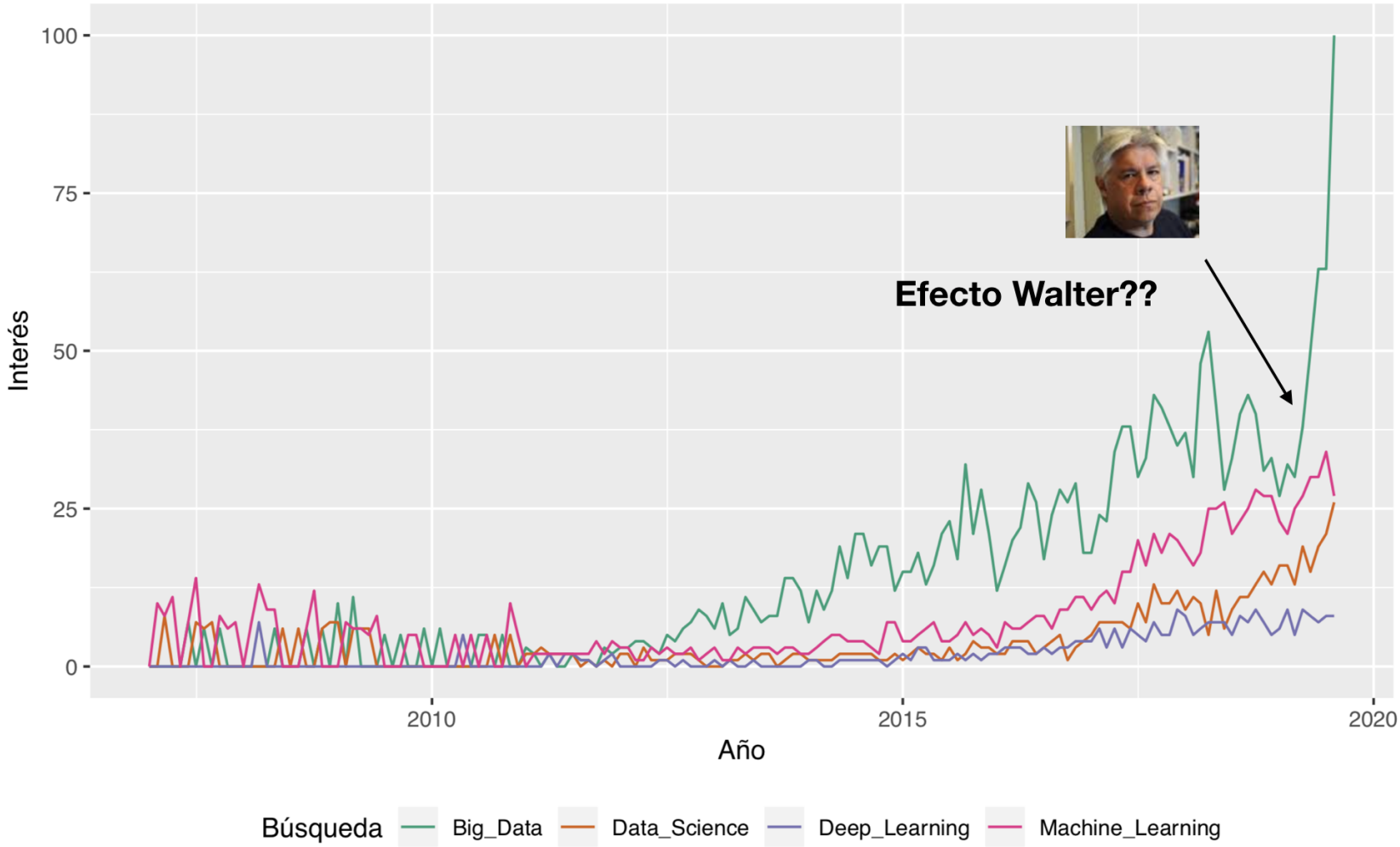
Google trends, mundo



Google trends, Argentina



Google trends, Argentina



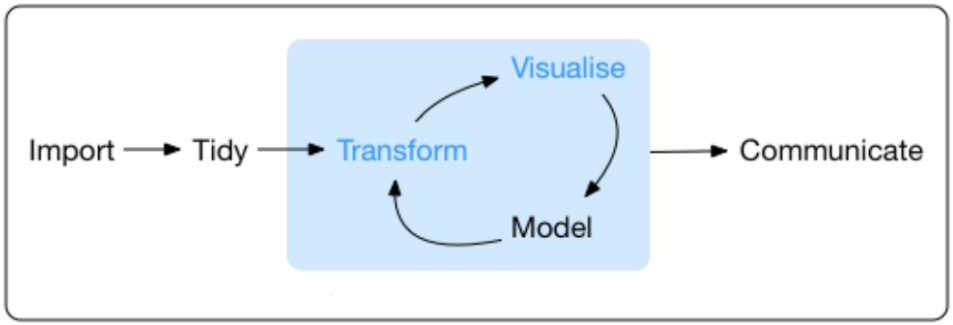
Denominador común

¿Cuál es el denominador común en todos esos términos?

Denominador común

¿Cuál es el denominador común en todos esos términos?

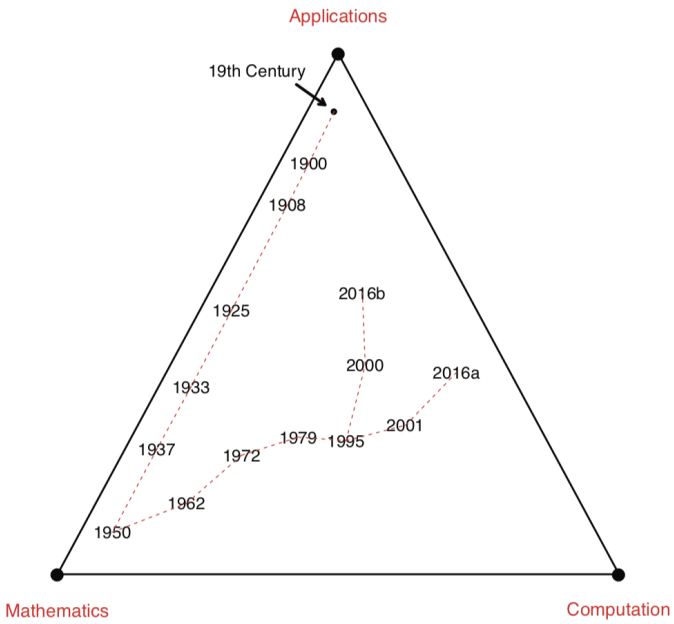
Estadística es una ciencia transversal que se encarga de recolectar información, analizar y entender los datos y modelar la incertidumbre de los mismos.



Estadística

- El centro del campo de la estadística se ha movido en los últimos 60 años desde lo más matemático y lógico a lo más computacional.
- Antes de la era de la computación (1950) era la era del cálculo y antes de big data trabajábamos con pequeñas muestras.

Evolución



Development of the statistics discipline since the end of the nineteenth century, as discussed in the text.

- 1900: Pearson: test χ^2
- 1908: Fisher: estadístico t-Student
- 1933: Pearson: test de hipótesis óptimo
- 1962: Tukey: el futuro del análisis estadístico
- 1963: Morgan y Sonquist: primer algoritmo de árboles
- 1979: Efron: Bootstrap
- 2001: Brieman: Random forest
- 2016: Ciencia de datos

Algoritmos e inferencia

"... algorithms are what statisticians do while inference says why they do them."

(Efron, B., Hastie, T. (2016))

Desafíos en la era de big data

La estadística es fundamental para asegurar la obtención de información precisa y con sentido de big data.

- Desarrollamos métodos estadísticos adecuados para big data
- Debemos fortalecer las habilidades de programación eficientes
- Herramientas que faciliten el limpiado y consistenciado de bases de datos
- Herramientas para almacenar grandes volúmenes de información

General

- Mayor **flexibilidad** en relaciones entre variables
- Aprendizaje supervisado o no supervisado

Supervisado: árboles de clasificación y regresión, bosques aleatorios, SVM, redes neuronales, ...

No Supervisado: análisis de grupos, componentes principales, MDS...

Explicar o predecir

Especial énfasis en performance predictiva:

1. Dividir los datos en entrenamiento y testeo.
2. Se evalúa la perfromance predictiva de los métodos en la muestra de testeo.
3. Se hace validación cruzada para seleccionar los parámetros del modelo o comparar modelos

Un modelo que predice muy mal fuera de la muestra, ¿puede dar explicaciones válidas y generalizables?

Tendencias recientes

- En economía, relación entre aprendizaje automático y causalidad.
- Se están desarrollando herramientas para que estos métodos sean más interpretables.
- Visualización estadística tanto en la exploración como para el diagnóstico de modelos.
- Otro desafío es la reproducibilidad.

Comentarios

- En economía y las ciencias sociales apostar más al trabajo colaborativo y grupos interdisciplinarios.
- Los nuevos desafíos implican el desarrollo de nuevas habilidades, incluir en los programas de maestría y doctorado estos temas.
- Aunque no se trabaje con aprendizaje automático incluir algunas de las enseñanzas, al menos separar en training y test.
- Un modelo que sólo explica la muestra no es generalizable.

GRACIAS

- Donoho, D. (2017). 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), 745-766.
- Efron, B., & Hastie, T. (2016). Computer age statistical inference (Vol. 5). Cambridge University Press
- Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 27(2), 83-85.
- Tukey, J. W. (1962). The future of data analysis. The annals of mathematical statistics, 33(1), 1-67.
- Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28.