¿Cómo auditar (parcialmente) los sistemas sociotécnicos?

Datos, ética y bien común. Medialab Prado

Javier Sánchez-Monedero
@javisamo
sanchez-monederoj at cardiff.ac.uk

23 marzo 2019

Cardiff University, UK

Data Justice Lab

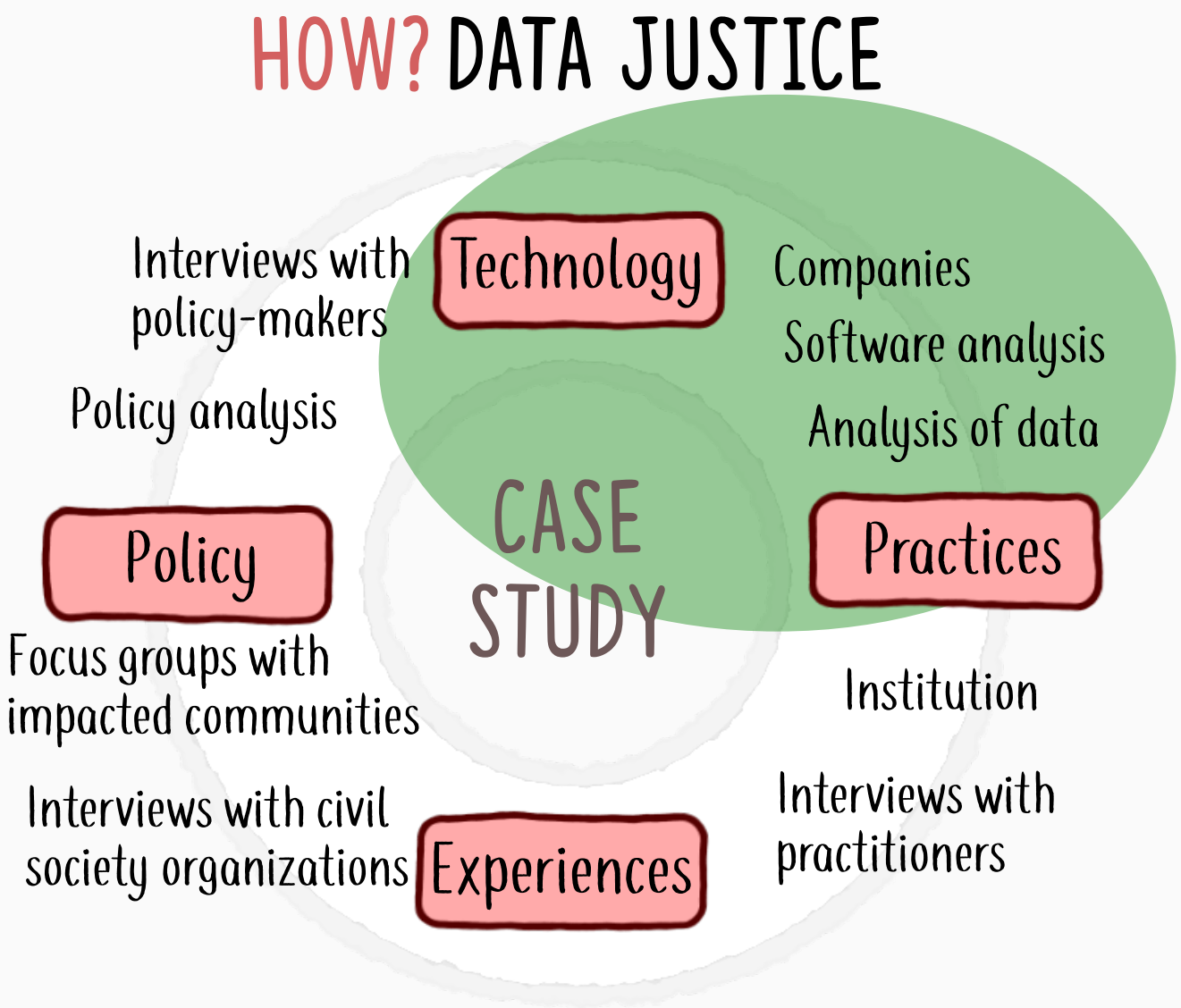
erc

datajusticelab.org

datajusticeproject.net

Data Justice

El marco 'Data Justice'



Ejes temáticos proyecto 'Data Justice'

Border control and migration

Law enforcement and policing

Low-wage work

Read more

Más en `https://datajusticeproject.net/`

Resumen rápido del aprendizaje máquina

Programación tradicional

Reglas explícitas:

```
si email contiene Viagra
  entonces marcarlo como
es-spam;
si email contiene ...;
si email contiene ...;
```

Ejemplos de Jason's Machine Learning 101

Programas de aprendizaje automático:

Aprender de los ejemplos:

```
intentar clasificar algunos
emails;
cambiar el modelo para
minimizar errores;
repetir;
```

...y luego utilizar el modelo aprendido para clasificar.

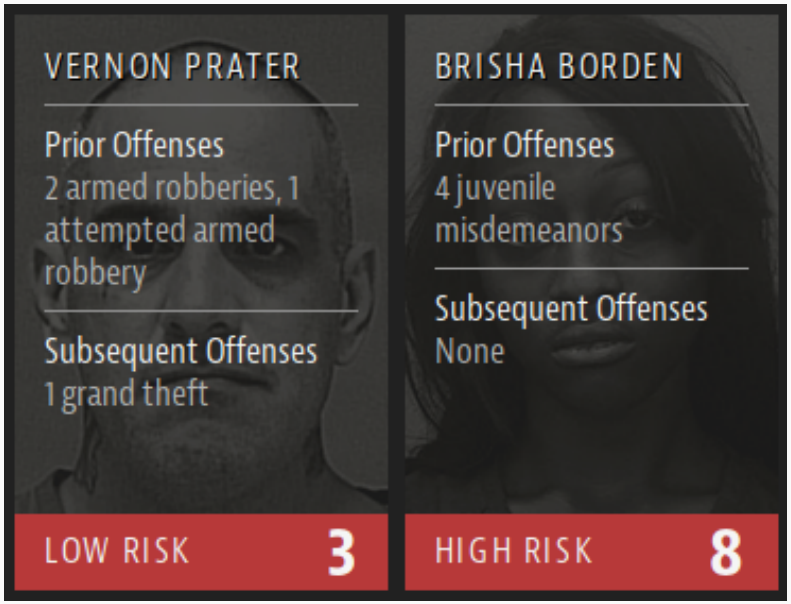
Como nadie está programando explícitamente a menudo se asume que es justo, no discrimina, está libre de sesgos humanos, etc. NOTA: además el código es por lo general difícil o imposible de auditar

El caso COMPAS

El caso COMPAS

- **COMPAS**: herramienta para calcular puntuaciones de riesgo de reincidencia de una persona en espera de juicio
- Utiliza ML para entrenar un modelo de estimación de **riesgo a partir de los registros históricos**
- **Variables de entrada**: historial criminal, tipo de cargos, género, grupo étnico, edad, *preguntas sobre el entorno*...
- **Variable dependiente**: grado de riesgo$\rightarrow$ los grados altos van a prisión preventiva

Discriminación racial



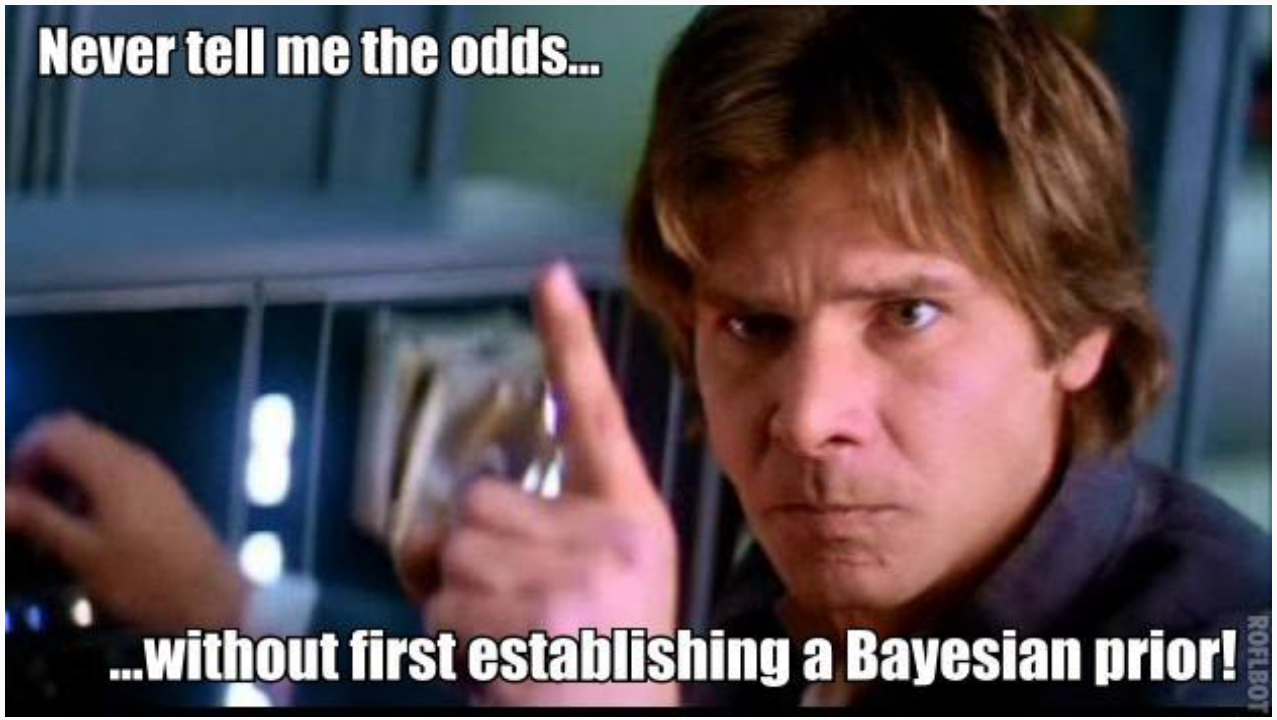
Fuente Angwin and Larson [2016]

ProPublica: el sistema discrimina porque sobrestima el riesgo para las personas afroamericanas (falsos positivos diferentes para los grupos: 44,8 % vd 23,4 %)

Northpointe: el sistema no discrimina porque clasifica el riesgo alto por igual (verdaderos positivos similares para todos los grupos étnicos: 63 % vs 59 %)

Be Bayesian

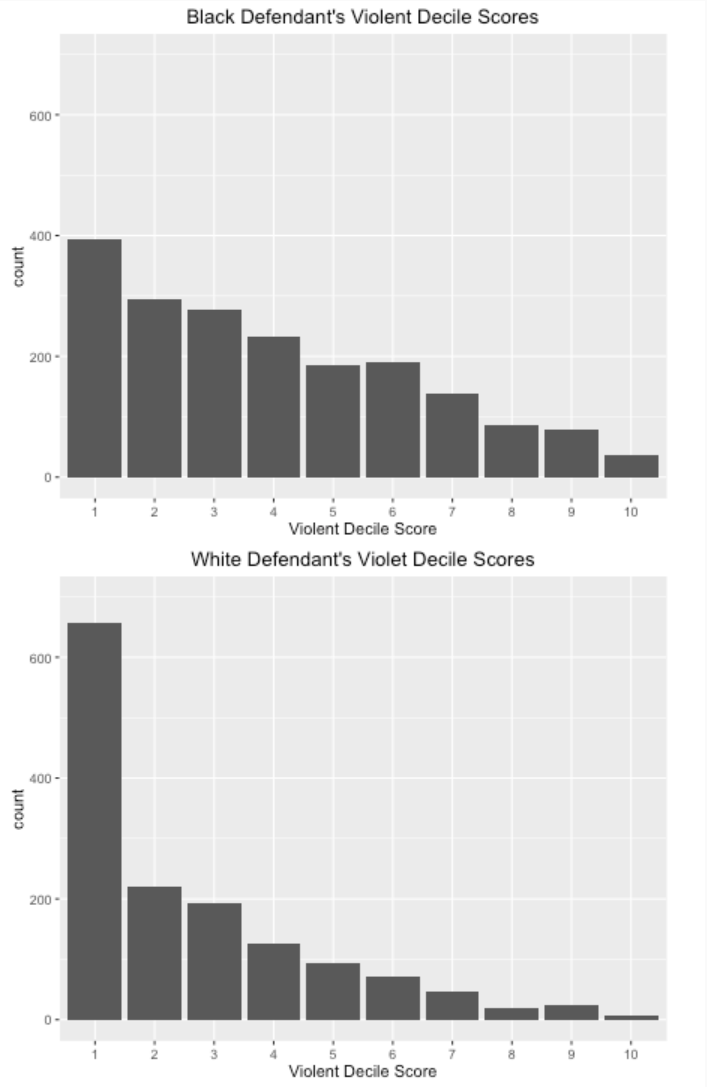
¿Cómo pueden ser compatibles las definiciones matemáticas de equanimidad de ProPublica y Northpointe?



Fuente Han Solo and Bayesian Priors

Be Bayesian II

Es matemáticamente compatible porque la prevalencia/frecuencia base/probabilidad a priori de los dos grupos es diferente Chouldechova [2017].



Fuente Larson and Angwin [2016]

¿Y si quitamos la variable de 'raza'?

Table 1. Human versus COMPAS algorithmic predictions from 1000 defendants. Overall accuracy is specified as percent correct, AUC-ROC, and criterion sensitivity (d') and bias (β). See also Fig. 1.

	(A) Human (no race)	(B) Human (race)	(C) COMPAS
Accuracy (overall)	67.0%	66.5%	65.2%
AUC-ROC (overall)	0.71	0.71	0.70
d'/β (overall)	0.86/1.02	0.83/1.03	0.77/1.08
Accuracy (black)	68.2%	66.2%	64.9%
Accuracy (white)	67.6%	67.6%	65.7%
False positive (black)	37.1%	40.0%	40.4%
False positive (white)	27.2%	26.2%	25.4%
False negative (black)	29.2%	30.1%	30.9%
False negative (white)	40.3%	42.1%	47.9%

Fuente Dressel and Farid [2018]

. . . bueno, pero puede haber variables proxy hacia la variable 'raza'.

¿Y si quitamos casi todas las variables?

Table 2. Algorithmic predictions from 7214 defendants. Logistic regression with 7 features (A) (LR_7), logistic regression with 2 features (B) (LR_2), a nonlinear SVM with 7 features (C) (NL-SVM), and the commercial COMPAS software with 137 features (D) (COMPAS). The results in columns (A), (B), and (C) correspond to the average testing accuracy over 1000 random 80%/20% training/testing splits. The values in the square brackets correspond to the 95% bootstrapped [columns (A), (B), and (C)] and binomial [column (D)] confidence intervals.

	(A) LR_7	(B) LR_2	(C) NL-SVM	(D) COMPAS
Accuracy (overall)	66.6% [64.4, 68.9]	66.8% [64.3, 69.2]	65.2% [63.0, 67.2]	65.4% [64.3, 66.5]
Accuracy (black)	66.7% [63.6, 69.6]	66.7% [63.5, 69.2]	64.3% [61.1, 67.7]	63.8% [62.2, 65.4]
Accuracy (white)	66.0% [62.6, 69.6]	66.4% [62.6, 70.1]	65.3% [61.4, 69.0]	67.0% [65.1, 68.9]
False positive (black)	42.9% [37.7, 48.0]	45.6% [39.9, 51.1]	31.6% [26.4, 36.7]	44.8% [42.7, 46.9]
False positive (white)	25.3% [20.1, 30.2]	25.3% [20.6, 30.5]	20.5% [16.1, 25.0]	23.5% [20.7, 26.5]
False negative (black)	24.2% [20.1, 28.2]	21.6% [17.5, 25.9]	39.6% [34.2, 45.0]	28.0% [25.7, 30.3]
False negative (white)	47.3% [40.8, 54.0]	46.1% [40.0, 52.7]	56.6% [50.3, 63.5]	47.7% [45.2, 50.2]

Fuente Dressel and Farid [2018]

¡Incluso si sólo se usan las variables de edad y número total de condenas previas el sistema sigue sobre-estimando el riesgo para la comunidad negra (columna B)!

El riesgo como proxy para la 'raza'

Reflexiones según Harcourt [2010]:

- Las herramientas de evaluación de riesgo/peligrosidad han ido reducciendo las variables predictivas y dando más importancia al historial criminal
- El riesgo queda ligado al historial criminal, y el historial criminal es un proxy para la raza
- En EEUU los intentos de utilizar métricas de peligrosidad han impactado negativamente en la comunidad afroamericana
- Las herramientas de evaluación de riesgo son una manera políticamente defendible de encarcelación masiva que proteje a los actores políticos

Problema del mundo real

¿Qué problema real está resolviendo el sistema? $\hat{y} = f(\mathbf{x})$

Respuesta de la prensa

- "Un algoritmo predice si los criminales volverán a delinquir"
- "Bienvenidos a Minority Report?"
- ...

Predicción real

El modelo de ML calcula un riesgo de reincidir-y-que-la-policía-le-pille

Verdad de fondo y bucles

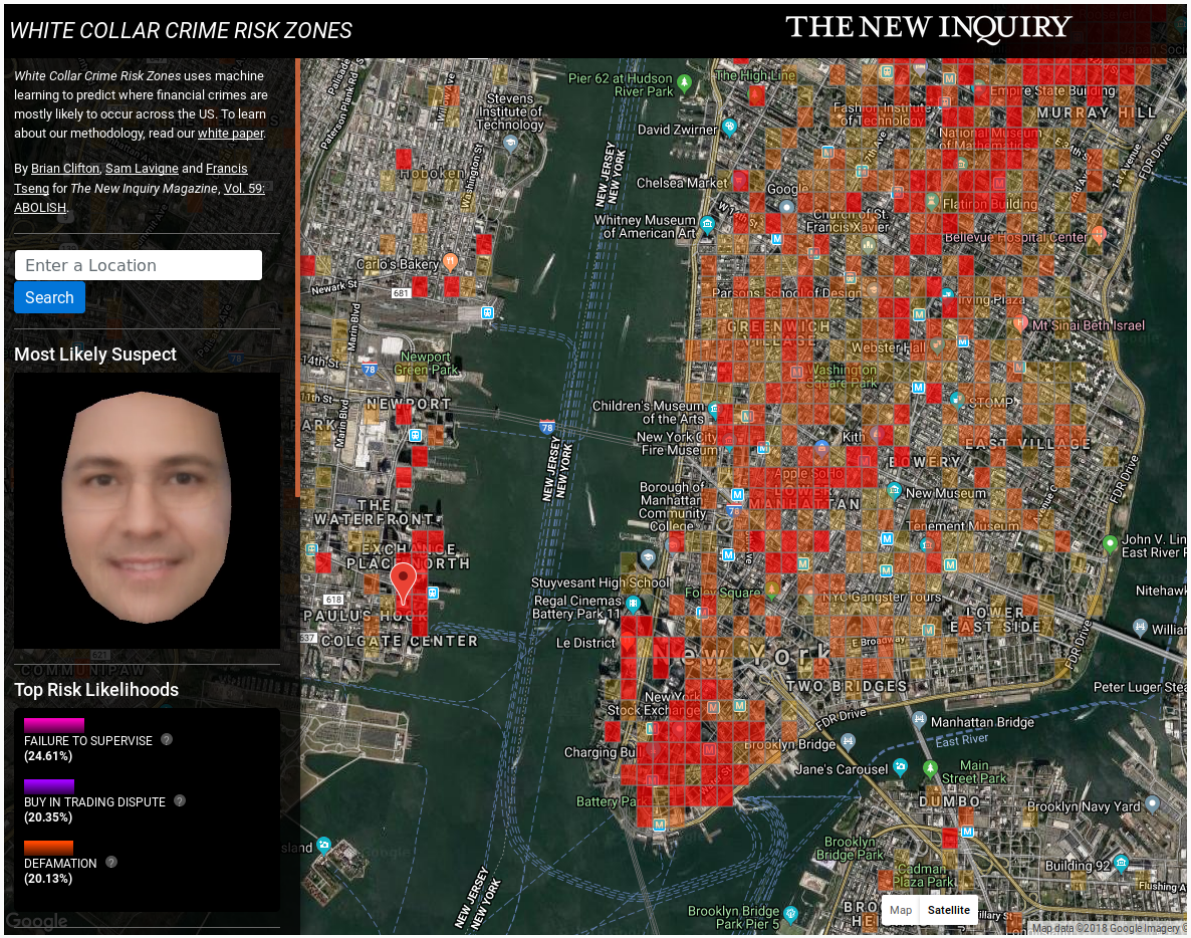
Fuente Lum and Isaac [2016]

Los **bucles de retroalimentación** pueden reproducir y amplificar los prejuicios Barocas and Hardt [2017], Ensign et al. [2017], ejemplo PredPol:

- No existe una base de datos de crímenes totales: sólo los denunciados o descubiertos
- La predicción de crimen en un área enviará recursos policiales a ese área
- Los eventos encontrados se añaden a la base de datos
- Es menos probable que se observen eventos que contradigan las predicciones

CODA: detrás de los números

¿Arreglar los problemas de sesgo para democratizar la vigilancia? ¿Policía predictiva o hay algo más?



whitecollar.thenewinquiry.com

Debate

Repaso

Preguntas:

- Cuáles son las variables y qué representan
- Cuál es la tarea que resolvemos en el mundo real
- ¿Se puede definir matemáticamente la ecuanimidad y justicia?
- ¿Clasificar, predecir, asignar nivel?
- ¿Cómo haríamos un análisis similar en Europa, donde no se recoge (o hasta está prohibido) el grupo étnico?
- ¿Lo nuevo funciona mejor o peor que lo anterior?

Working Paper

How to (partially) evaluate automated decision systems. Working paper by Javier Sánchez-Monedero and Lina Dencik. December 2018. https://datajusticeproject.net/working-papers/

¿Preguntas?¡Gracias!

elsaltodiario.com/post-apocalipsis-nau

Referencias i

J. Angwin and J. Larson. Machine Bias. *ProPublica*, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

S. Barocas and M. Hardt. Fairness in Machine Learning. NIPS 2017 Tutorial, 2017. URL http://fairml.how/.

A. Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, June 2017. ISSN 2167-6461, 2167-647X. doi: 10.1089/big.2016.0047. URL http://www.liebertpub.com/doi/10.1089/big.2016.0047.

J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1): eaao5580, Jan. 2018. ISSN 2375-2548. doi: 10.1126/sciadv.aao5580. URL http://advances.sciencemag.org/content/4/1/eaao5580.

D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway Feedback Loops in Predictive Policing. *arXiv:1706.09847 [cs, stat]*, June 2017. URL http://arxiv.org/abs/1706.09847. arXiv: 1706.09847.

B. E. Harcourt. Risk as a Proxy for Race. SSRN Scholarly Paper ID 1677654, Social Science Research Network, Rochester, NY, Sept. 2010. URL https://papers.ssrn.com/abstract=1677654.

J. Larson and J. Angwin. How We Analyzed the COMPAS Recidivism Algorithm, May 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5):14–19, Oct. 2016. ISSN 17409705. doi: 10.1111/j.1740-9713.2016.00960.x. URL http://doi.wiley.com/10.1111/j.1740-9713.2016.00960.x.