Los grandes problemas de "Big Data"

Jeanna Matthews

Spring School on Networks

Redes para ciudades inteligentes

20 octubre 2017

Data&Society

datasociety.net

Un poco sobre mí

- Profesora de computacion en la Universidad de Clarkson
- Fellow en Data and Society
- Co-presidente de un ACM Subcomité de Responsabilidad y Transparencia Algorítmico

Los grandes problemas de "Big Data"

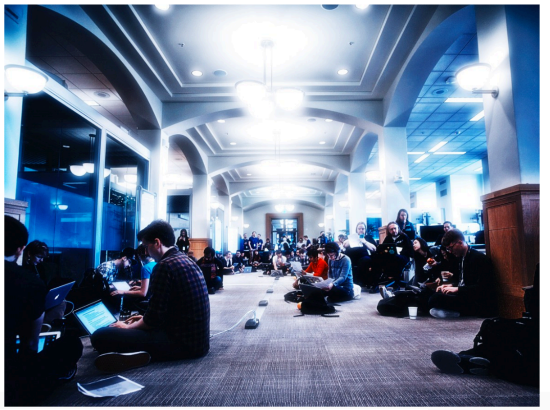
- Falta de Privacidad / espejismo de privacidad
- Espejismo de anonimato
- Decisiones grandes/ decisiones injustas

El poder de “Big Data”

- Comprender / optimizar nosotros mismos y el mundo
  - Mejorar nuestras ciudades, curar enfermedades, usar recursos limitados de una manera efectiva, …
- Mucho poder para el bueno . y el malo
- Redes para ciudades inteligentes *y* para la sociedad en que queremos vivir?

MEGAN MOLTENI SCIENCE 02.13.17 5:35 PM

DIEHARD CODERS JUST RESCUED NASA'S EARTH SCIENCE DATA



Human Face of Big Data

Big Data


EACH OF US NOW LEAVES A TRAIL OF DIGITAL EXHAUST, AN INFINITE STREAM OF PHONE RECORDS, TEXTS, BROWSER HISTORIES, GPS DATA, AND OTHER INFORMATION, THAT WILL LIVE ON FOREVER.
Instead of "find my iPhone," some auto insurance companies are offering a service that may enable parents to "find my teenager." Progressive Insurance, for example, offers the Snapshot, a tracking device that reports on a car's location, acceleration, braking, and distance traveled. Owners who install the device can get a 10 to 15 percent discount on their policy. Privacy activists, however, fear the technology is ripe for abuse.

Nuestro escape digital

- Emails, texts
- Social media
- Web browsing history, web site use and cross site correlations
- Cell phone location
- Purchase history, credit cards, wish lists, products viewed/ reviewed, frequent buyer cards,
- Cameras (yours, others, on street, accidental, aware/ unaware, facial recognition) , GPS tags in pictures
- Fitbits, microphones, Google glass,
- License plate readers, passport use, radio-frequency identification (RFID) readers, satellite imagery
- E-readers, streaming video use, MOOCs,

Dispositivos como Fitbit

- Fitbit
- Patrones de actividad,
- Patrones de sueño
- Ritmo cardíaco

- ¿Promesa? Ejemplo: Más datos sobre patrones de sueño que nunca antes en la historia
- ¿Problema? Dos personas en el mismo lugar con alta frecuencia cardíaca?
- ¿Sería difícil hacer un software que nos permita ver esos datos nosotros mismos y no entregarlos a la nube?

MOOCs

- Massively Open Online Courses
- ¿Cuántas veces se hizo una prueba o intentó un ejercicio? ¿Cuánto tiempo pasaste leyendo? ¿Has leído todas las páginas?

- ¿Promesa? Experimentos dirigidos sobre la mejor manera de enseñar en una escala masiva: ¡más datos que nunca!
- ¿Problema? Vender a los empleadores? ¿Quién es bueno en esta tarea? Reputación que comparte si participas o no

¿A dónde van los datos?

- Uso interno
- Compartido con socios de “confianza”
  - Otros negocios? El gobierno? investigadores como nosotros? :-
- Correlaciones entre conjuntos de datos
- Compartido "anónimamente" pero posible para deanomizar
- Robado por hackers
- Abuso de información privilegiada?

Datos? Predicciones?

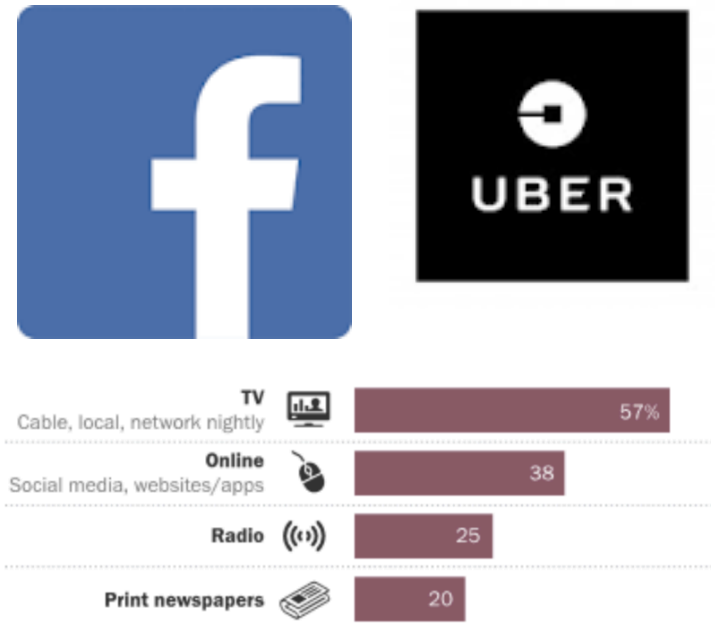
- Una cosa son los datos que son verdaderos
  - ¿Sabías que estabas revelando? ¿No lo sabías? correlaciones?
  - Suficientemente malo?
- Datos Incorrectos?
- Predicciones?
  - Haga clic en "Me gusta" en la página de Facebook para una enfermedad -> tienes esa enfermedad
  - Compre vitaminas prenatales o frecuencia cardíaca en Fitbit -> estás embarazada

Usado cada vez más para grandes decisiones

- Todos hemos visto esto usado para publicidad
  - Busque algo en Google y vea anuncios en Facebook y Amazon
- Usado en algoritmos de aprendizaje automático
  - Predicciones: decisiones de crédito, fianzas, admisiones universitarias, decisiones de vivienda, asignación de recursos públicos ....
  - ¿Mirar los atributos potencialmente discriminatorios o inferirlos?
- ¿Para qué más?

Algorithms and Platforms Reshaping Society

- Algoritmos entrenados de datos grandes utilizados cada vez más para decisiones de grandes vidas
  - Contratación, vivienda, vigilancia, recursos públicos, etc.
- Las plataformas como Facebook, Twitter y Uber tienen un profundo impacto en nuestras relaciones personales y públicas
  - ¿Cómo encontramos un trabajo?
  - ¿Cómo recibimos nuestras noticias?
  - ¿Cómo encontramos un esposo?



Connectivity

First Evidence That Online Dating Is Changing the Nature of Society

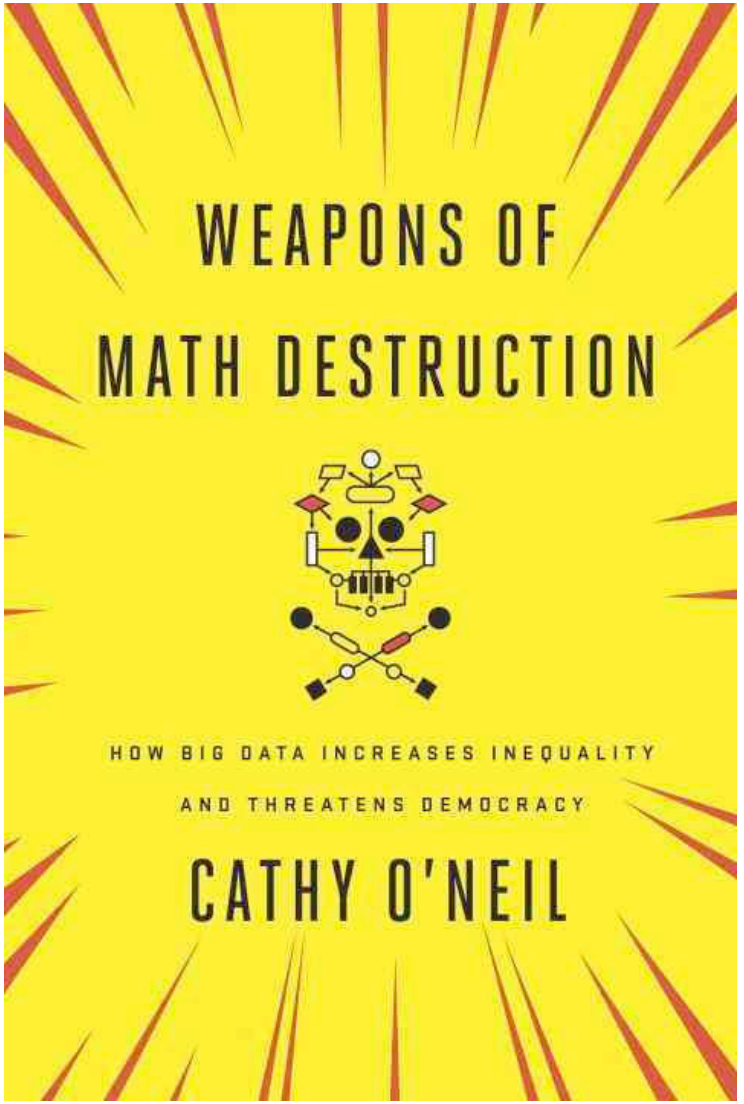
Dating websites have changed the way couples meet. Now evidence is emerging that this change is influencing levels of interracial marriage and even the stability of marriage itself.

by Emerging Technology from the arXiv October 10, 2017

Ejemplos

- ¿Donde trabajas?
  - ¿La ordenación de currículos para trabajos? ¿Le mostrarán anuncios de trabajos apropiados? ¿Cómo se juzga tu trabajo?
- ¿Donde vives?
  - ¿Puedes comprar una casa? ¿Tendrás acceso al crédito? ¿Le mostrarán anuncios de casas apropiadas?
- ¿Cómo se toman estas decisiones?
  - ¿Qué derechos tenemos para entender los prejuicios construidos por los programadores o los sesgos más probables aprendidos de los datos históricos?

Weapons of Math Destruction por Cathy O'Neil

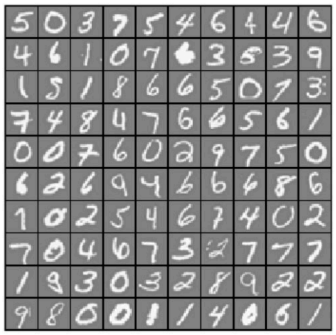


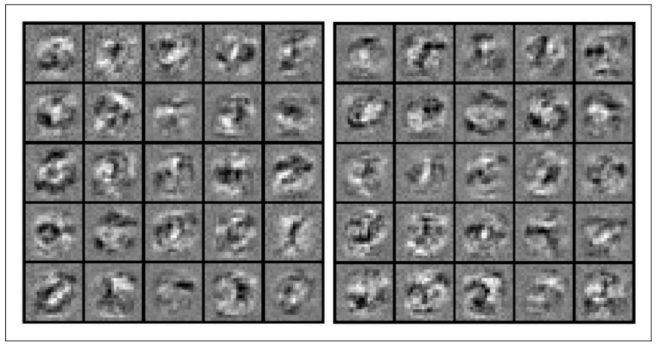
¿Cuáles son los objetivos? ¿Quién es el cliente?

- Algoritmos optimizados para la eficiencia / riesgo reducido para el decisor
  - Protección para las personas afectadas por las decisiones?
- Plataformas optimizadas para publicidad
  - Protección para la sociedad? ¿Democracia?
  - Actores que deliberadamente "gamifican" el sistema
- Necesitamos examinar activamente estos algoritmos y estas plataformas en beneficio de las personas y la sociedad.

En la caja negra

- Aprendizaje automático de dígitos
- Aprendizaje automático de lobos y perros
- Aprendizaje automático de nuestro escape digital?
  - Credit card charge for marriage counseling => raise interest rates, lower credit limit
  - Credit card charge? Facebook like? Colleague on Linked In? Happen to be near a demonstration site?





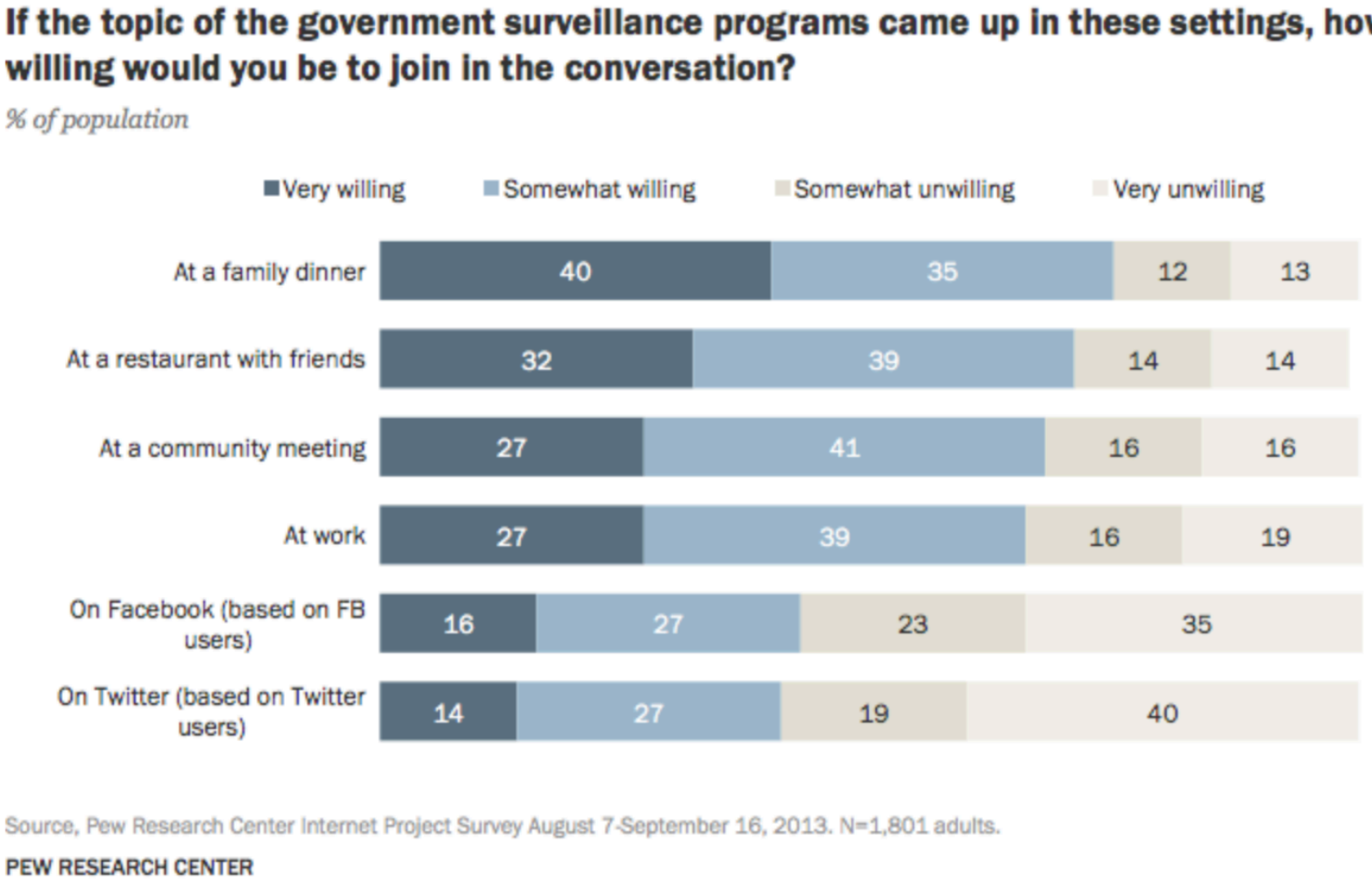
(a) Husky classified as wolf (b) Explanation

Figures from “How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms”, Burrell and ““Why Should I Trust You?": Explaining the Predictions of Any Classifier”, Ribeiro et al.

Algoritmos propietarios

- Algoritmos propietarios utilizados en decisiones públicas
- Ejemplo en los EE. UU. de software COMPAS utilizado para asignar una puntuación numérica a la probabilidad de una persona de cometer otro delito
  - Caso Loomis vs. Wisconsin
  - Proteger la propiedad intelectual de la empresa sobre los derechos de los demandados para comprender cómo se calcula su puntaje
- Codigo abierto? ¿Conjuntos de datos de entrenamiento? Conjuntos de reglas en constante evolución?

La incapacidad para predecir el costo de su acción conduce a un efecto de enfriamiento en el discurso civil



Man is to Computer Programmer as Woman is to Homemaker?

- Entrenados en textos de calidad relativamente alta como los artículos de Google News (no texto deliberadamente sesgado) muestran fuertes estereotipos
- Stilizados en innumerables aplicaciones desde la búsqueda web hasta la ordenación de currículos para trabajos

Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

¿Decisiones sin prejucios tomado por computadora?

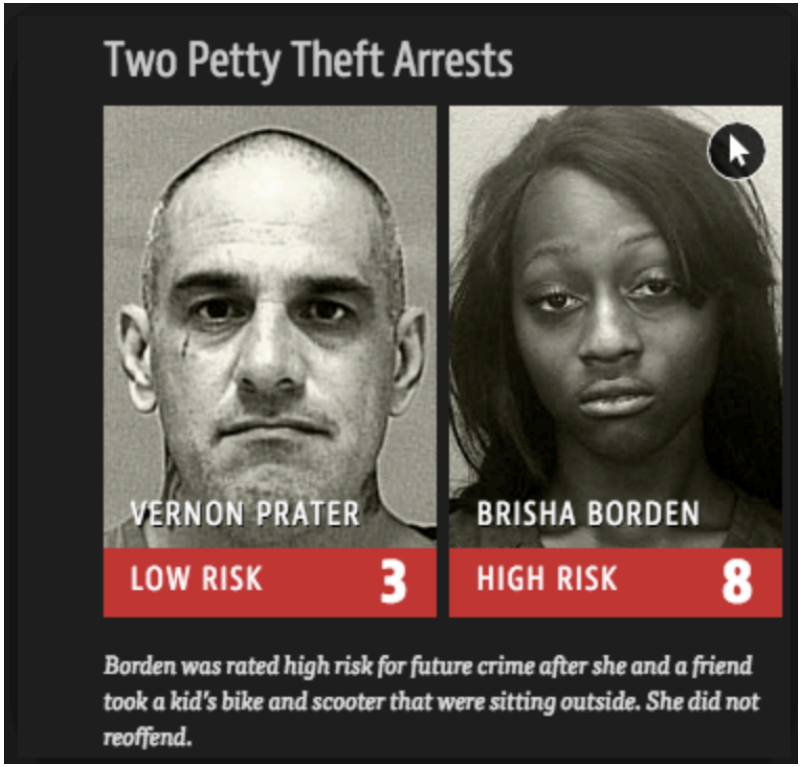
- Intento deliberado de llamar decisiones lógicas / pura de prejuicios humanos
- ncluso puede ignorar deliberadamente atributos / características como la raza, el sexo o la religión, pero muchos otros "proxies" para esta información directa
  - Dirección -> raza?
  - Hábitos de compra -> género?

Ejemplos de “proxies”

- Algunos descriptores aparentemente imparciales, operan como referentes de raza
  - Zip code, ancestry, disease predisposition, linguistic characteristics, last name, criminal record, and socioeconomic status
- Para el género? Para la religión? Para el estado de discapacidad?

Eficiencia y equidad?

- Eficiencia / utilidad de la toma de decisiones para el decisor
  - Algunos atributos sensibles pueden ser predictivos
  - ¿Podemos usarlos? Proxies para ellos?? Proxies for them?
- No podemos tenerlo todo!



Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg * Sendhil Mullainathan † Manish Raghavan ‡

Algorítmicos de Transparencia y Responsabilidad

- US-ACM/EUACM Statement on Algorithmic Transparency and Accountability
  - Un documento corto
  - 7 principios
  - En ingles: Awareness, access and redress, accountability, explanation, data provenance, auditability, validation and testing
- Un articulo en CACM "Toward Algorithmic Transparency and Accountability"

https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

Bottom Line

- Nuestros datos, los algoritmos y las plataformas no son usados * solo * sugestivos y entretenidos, no solo usados en optimizaciones anónimas de alto nivel
- Cada vez más utilizado para grandes decisiones sobre la vida de cada personas y fundamentalmente están cambiando nuestra sociedad
- Para construir el mundo que queremos, necesitamos estos algoritmos y plataformas para ser responsables y transparentes