Límites en la minería de trazas digitales

Daniel Gayo Avello

Última modificación: Mon, 26 Sep 2022 10:58:20 GMT

Tabla de contenidos

Introducción

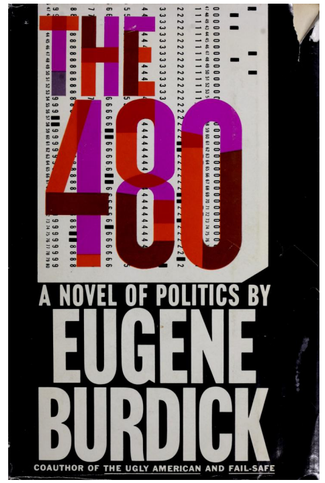
This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. ***Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.***

Introducción

If you had to dream of research content, it would be sending out a ***diary*** *and having people* ***record their thoughts at the moment****. That's like a* ***social scientist's wet dream****, right? And here it has kind of fallen on our lap,* ***these ephemeral recordings that we would not have otherwise gotten****.*

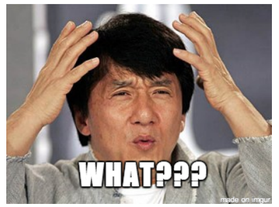
Introducción

The new underworld is made up of innocent and well-intentioned people who work with slide rules and calculating machines and computers which can retain an almost infinite number of bits of information as well as sort, categorize, and reproduce this information at the press of a button. Most of these people are highly educated, many of them are Ph.D.s, and none that I have met have malignant political designs on the American public. They may, however, radically reconstruct the American political system, build a new politics, and even modify revered and venerable American institutions-facts of which they are blissfully innocent. They are technicians and artists; all of them want, desperately, to be scientists.



Introducción

- Ya mencionamos que los datos generados por usuarios se han considerado una oportunidad única para entender mejor el funcionamiento de la sociedad moderna e incluso conocer mejor a los individuos. Ejemplos:
  - Detectar y monitorizar epidemias [1] [2] [3] [4] [5] [6]
  - Entender mejor los problemas de salud [7] [8] [9] [10] y salud mental [11] [12] [13] [14] [15].
  - Conocer los efectos secundarios de ciertos medicamentos [16] [17] [18].
  - Cuantificar los patrones de consumo de comida [19] [20] [21] [22] [23]
  - Pulsar la **opinión** [24] 😡 [25] 😠 [26] 🤬 [27] 😒 y las **emociones** del público [28][29][30].
  - Predecir para usuarios individuales la **orientación sexual, etnicidad, punto de vista religioso y político, rasgos de personalidad, inteligencia, felicidad, uso de sustancias adictivas, separación parental, edad y género** [31][32][33]
  - Utilizar los medios sociales como fuente de los servicios de inteligencia (aka SOCMINT) [34] [35]
  - **¡Entrenar modelos que pueden emitir juicios morales!** 😳 *Delphi* [37]
- La conveniencia de muchas de esas aplicaciones es, como poco, discutible...
- También se dijo que esto puede funcionar... [37] [38] [39] [40] [41] [42]



Introducción

- Una primera cuestión a considerar es que las **trazas digitales** deben manipularse con precaución puesto que la mayor parte no son más que *digital exhaust*, un subproducto que nunca se diseñó para su consumo directo... (véase una perspectiva de los datos como **residuo tóxico**).
- De hecho, ya se señaló cómo las consultas recomendadas por un buscador pueden llevar a alguien a una postura extrema:

```
vaccines
vaccines pros and cons
why vaccines are bad | disadvantages of vaccines | vaccines to avoid
```

- El problema es que para (casi) cualquier consulta se sugieren al usuario formas de completarla según la va formulando...

  "por qué" (según Google, Bing y Yahoo!), **"para qué"** (según Google, Bing y Yahoo!), **"cuándo"** (según Google, Bing y Yahoo!), **"quién"** (según Google, Bing y Yahoo!), **"cuánto"** (según Google, Bing y Yahoo!), **"qué hacer cuando"** (según Google y Yahoo!), **"qué hacer para"** (según Google, Bing y Yahoo!), **"es bueno"** (según Google, Bing y Yahoo!), **"es malo"** (según Google, Bing y Yahoo!), **"son malos"** (según Google, Bing y Yahoo!), **"son malas"** (según Google, Bing y Yahoo!), **"es peligroso"** (según Google y Yahoo!), **"son peligrosos"** (según Google, Bing y Yahoo!), **"son peligrosas"** (según Google, Bing y Yahoo!), **"como conseguir"** (según Google, Bing y Yahoo!), **"dónde conseguir"** (según Google, Bing y Yahoo!), **"dónde comprar"** (según Google, Bing y Yahoo!), **"cómo hacer"** (según Google, Bing y Yahoo!), **"cómo se hace"** (según Google, Bing y Yahoo!), **"cómo puedo fabricar"** (según Google y Yahoo!), **"los profesores"** (según Google, Bing y Yahoo!), **"los políticos"** (según Google, Bing y Yahoo!), **"ciudadanos"** (según Google y Bing), **"podemos"** (según Google), **"psoe"** (según Google y Bing), **"vox"** (según Google y Bing), **"los asturianos son"** (según Google y Bing), **"las asturianas son"** (según Google), **"los inmigrantes"** (según Bing y Yahoo!).

- ...completarla con una buena porción de **sesgos y prejuicios aportados por los usuarios** del buscador.
- Una herramienta que os puede resultar llamativa es *Web Seer*.
- Además, esos datos pueden malinterpretarse: *Corona is the world's most googled beer* 😂. *Google Trends* para cervezas, Budweiser (cerveza) vs "budweiser" (término), Corona (cerveza) vs "corona" (término), Einstein (físico) vs "albert einstein" (término).

Introducción

- Esos ejemplos son meras anécdotas y la presencia de **sesgos** (o la **falta de representatividad** de los usuarios frente a la población general) son solo algunos de los muchos problemas que plantea el análisis de trazas digitales.
- En consecuencia:
  - Debemos reconocer las **limitaciones** de las trazas digitales.
  - Debemos reflexionar sobre las **consecuencias** de estas técnicas y sobre su **conveniencia**. Dicho de otra manera, hay que considerar las **implicaciones éticas**.
  - Debemos ser conscientes de que la creencia acerca de la precisión y objetividad de estos métodos es solo eso: una creencia.
- **A todo eso dedicaremos esta unidad.**

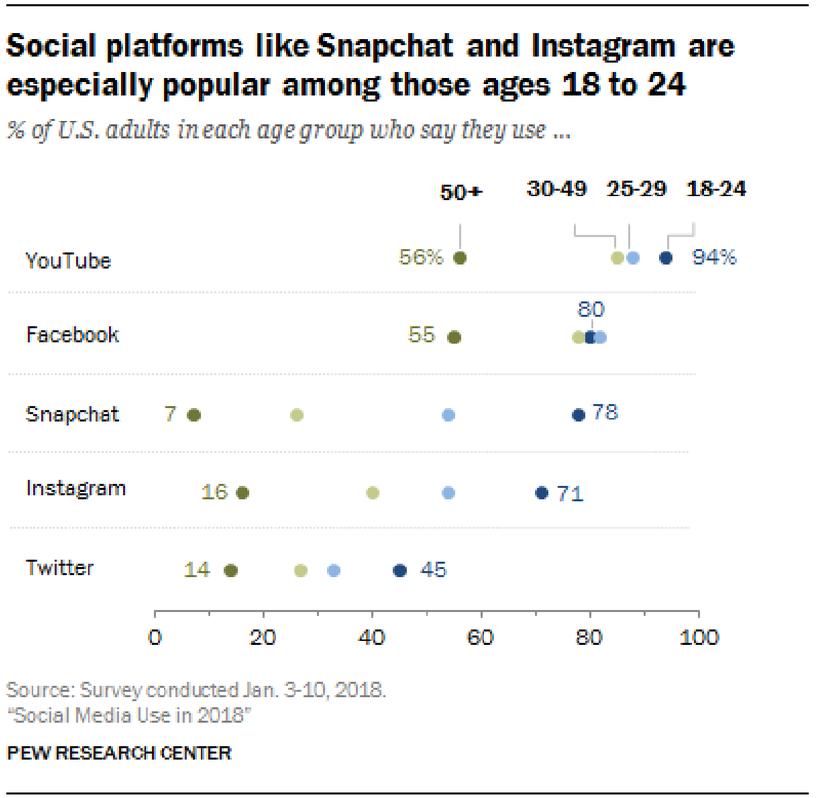
Limitaciones de las trazas digitales

Big Data presents ***new opportunities for understanding social practice****. Of course the next statement must begin with a* ***"but."*** *And that "but" is simple:* ***Just because you see traces of data doesn't mean you always know the intention or cultural logic behind them. And just because you have a big N doesn't mean that it's representative or generalizable.***

Limitaciones de las trazas digitales

Sesgos demográficos

- La composición demográfica de las distintas plataformas es muy diferente de la de la sociedad (aka, la base de usuarios **no** es una muestra representativa).
- Tomemos como ejemplo un estudio realizado en EE.UU. en 2018:
  - No hay dos plataformas que tengan la misma pirámide demográfica.
  - Algunas plataformas capturan a casi todos los miembros de algunos estrato de edad pero no a todos (y aún así puede haber sesgos).
  - En varias plataformas hay una **sobrerrepresentación de los estratos más jóvenes** (dicho de otra manera, sesgo).
  - **Twitter**, la plataforma más habitualmente usada a modo de termómetro de la opinión pública, es con diferencia **la que menos penetración tiene** (dicho de otra manera, no se usan los mejores datos sino los más fáciles de obtener).
  - Entre las distintas plataformas hay **diferencias marcadas por género** (p.ej., el 41% de las mujeres estadounidenses usan Pinterest frente al 16% de los hombres), por **extracción socioeconómica** (p.ej., el 50% de los estadounidenses con formación universitaria usan LinkedIn frente al 9% de los que no la tienen), por **etnicidad** (el 49% de los hispanos de EE.UU. usan WhatsApp frente al 21% de los negros y el 14% de los blancos) y por **lugar de residencia** (todas las plataformas tienen mucha mayor penetración en ciudades que en entornos rurales). [+]



Limitaciones de las trazas digitales

Sesgos demográficos

- Esa falta de representatividad no es reciente ni afecta a las plataformas actuales únicamente. Siempre ha estado ahí y afecta a cualquier plataforma digital: p.ej., Facebook, MySpace, Xanga, and Friendster en 2007 o Twitter en 2011.
- Por otro lado, **no todos los grupos demográficos utilizan las plataformas de la misma manera**:
  - Por ejemplo, se han encontrado diferencias entre **hombres y mujeres** en el uso de características de LinkedIn [1].
  - **Estudiantes y profesorado** universitario también se diferencian en el uso que hacen de distintos medios sociales [2].
  - **Distintas nacionalidades** hacen uso en distinta medida de características de Twitter como retuits, *replies*, menciones, hashtags y URLs [3].
  - Usuarios de **distintas edades** suben fotos, hacen comentarios y usan el chat de Facebook con frecuencias muy diferentes [4].
- Además, **distintas plataformas se usan para diferentes necesidades:**
  - Por ejemplo, si alguien busca interacción social y afecto usa Facebook mientras que si quiere dejar constancia de su descontento utiliza un foro [5]
  - Incluso plataformas usadas fundamentalmente para interacción social como Facebook y mensajería instantánea se emplean de forma diferente, p.ej. para una comunicación más íntima y profunda los usuarios prefieren la mensajería sobre Facebook [6]
- Por último, y para empeorar aún más la situación, **todas estas diferencias cambian a lo largo del tiempo de manera impredecible**; es decir, la composición demográfica de los usuarios [7] [8] que participan en un momento dado y sus acciones [9] [10] pueden ser radicalmente diferentes a las de otro momento.
- En resumen, **las trazas digitales que dejan (o no) los usuarios varían según el estrato demográfico y con el tiempo y, además, la base de usuarios no solo no es representativa de la sociedad sino que esa falta de representatividad muta continuamente.**

Limitaciones de las trazas digitales

Sesgos de autoselección

- Además del problema de la falta de representatividad las trazas digitales exhiben un **sesgo de autoselección**. Es decir, tenemos los rastros que han dejado aquellos que querían dejar un rastro: no se puede obligar a nadie a hacer una consulta en la Web, comentar en un foro, retuitear o hacer un *like*.
- Los principales problemas derivados de la autoselección son los siguientes:
  - La inhibición de los usuarios no es un fenómeno minoritario, de hecho, **los usuarios "silenciosos" son una parte muy importante de las comunidades online** [1] **[2]**
  - Los usuarios silenciosos tienen generalmente **opiniones diferentes** de aquellos que sí están dejando trazas digitales **[2]**[3] y tienden a ser más **moderados** [4]
  - Los **usuarios que sí dejan trazas** no solo difieren en opinión de los silenciosos sino que tienden a ser mucho **más activos** llevando a la impresión de que su postura es mayoritaria **[2]** y también **más extremos en sus opiniones** [5]
  - Todo esto puede generar una **espiral de silencio** donde la minoría vocal se activa cada vez más y la mayoría silenciosa se desmoviliza aún más [6]

Limitaciones de las trazas digitales

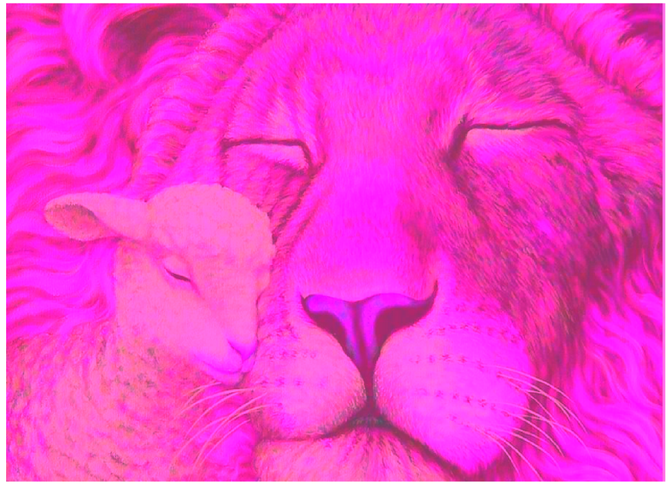
Sesgos en la producción de contenidos

- Los usuarios no publican por igual acerca de cualquier tipo de contenido sino que **la producción de contenidos también está sesgada**. Ejemplos:
  - Se producen **más contenidos para eventos más recientes** que más antiguos, p.ej. en un estudio de 2003 la última década era responsable del 85% de los resultados. La Wikipedia muestra un sesgo semejante hacia eventos recientes [1]
  - **Los eventos y sentimientos más extremos dan lugar a más contenidos** que los no extremos [2] [3]
  - **Los sentimientos positivos generan más contenidos** que los sentimientos negativos [3]
  - **Las experiencias más inusuales están sobrerrepresentadas** en la Web mientras que las experiencias más comunes están infrarrepresentadas [4]
  - **Los contenidos falsos se expanden más** (llegan a más gente) **y más rápidamente** que los contenidos auténticos [5]
- Todo esto quiere decir que, **dependiendo de la naturaleza del fenómeno que se quiera estudiar, éste puede estar sobrerrepresentado o infrarrepresentado** y, además, no hay forma de saber *a priori* en qué sentido irá el sesgo.

Limitaciones de las trazas digitales

¿Sesgos del mundo *offline*?

"El lobo y el cordero serán apacentados juntos, y el león comerá paja como el buey; y el polvo será el alimento de la serpiente."



Limitaciones de las trazas digitales

Sesgos del mundo *offline* (aka la cruda realidad)

- El **sexismo** *online* es rampante:
  - Los hombres promocionan a otros hombres a expensas de las mujeres mientras que las mujeres promocionan más a los hombres que a otras mujeres [1]
  - Las mujeres tienden a ser cosificadas y a recibir mensajes fuera de tópico y sexualizados [2]
  - La cultura de la violación está muy presente en los foros de discusión [3]
  - En ocasiones los usuarios varones orquestan campañas de odio y acoso a mujeres [4]
- El **lenguaje homofóbico** es extremadamente frecuente (véase NoHomophobes.com).
- El **racismo** *online* es habitual [5][6][7]
- Esas actitudes **silencian o expulsan** a parte de los usuarios, sesgando aún más los datos hacia las posiciones de quienes discriminan y oprimen.
- Además, **todos estos sesgos, prejuicios y actitudes discriminatorias y derogatorias que ocurren *online* se transfieren a los *datasets*** [8] que se usan para **entrenar sistemas "inteligentes"** que, a su vez, se despliegan *online* y **que discriminan de manera automatizada reforzando la discriminación ya existente, reproducen el discurso del odio** (p.ej., aquí, aquí o aquí) o, cuando menos, **ofrecen información inexacta como si fuera fidedigna**.
- Ni que decir tiene, se insiste en dar solución técnica a un problema social (ejemplo).

Limitaciones de las trazas digitales

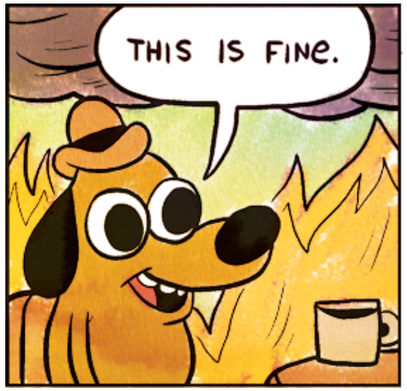
Entorno adversarial

- A lo largo de la asignatura se ha insistido en que Web es un entorno adversarial desde sus inicios.
- Las bases de usuarios no solo no son representativas de la sociedad sino que, además, abundan las cuentas automatizadas total o parcialmente que fingen ser personas reales.
- Además de los sesgos de autoselección y de producción de contenidos va a haber información fabricada con la intención de intoxicar.
- Es decir, **los usuarios maliciosos y la información falsa pueden adulterar irremediablemente los datos con los que se pretenda hacer minería web, los sistemas inteligentes que se entrenen con esos datos y falsear las conclusiones que se obtengan.**
- Más información en el tema anterior "Comportamiento malicioso en la Web".

Limitaciones de las trazas digitales

En pocas palabras...

- Las plataformas sociales exhiben **sesgos demográficos** manifiestos en sus bases de usuarios y **no son muestras representativas de la sociedad.**
- Distintos estratos demográficos usan distintas características de las plataformas de diferentes formas y con diferentes finalidades.
- Todo esto **cambia con el tiempo de manera impredecible.**
- Las plataformas sociales exhiben **sesgos de autoselección** claros que, además, llevan a silenciar las posturas más moderadas y reforzar las extremas.
- Existen también **sesgos en la producción de contenidos** sobrerrepresentando lo **reciente, inusual y extremo.**
- Además, **la información incorrecta y falsa se extiende más y con mayor rapidez.**
- Todos los **sesgos, prejuicios y actitudes discriminatorias y derogatorias** de la sociedad **se transfieren** a las plataformas *online*, a las trazas digitales que generan y, en consecuencia, **a los sistemas inteligentes** entrenados con ellas **que reproducen y refuerzan la discriminación pre-existente.**
- Por último, existen **usuarios maliciosos** que actuán con la única intención de **manipular** el ecosistema Web y, por ende, **alterar las conclusiones** que se puedan extraer de su estudio.



Ética en la minería de trazas digitales

Investigación académica

- En la introducción se mencionó investigación con trazas digitales que resulta, como poco, controvertida; por ejemplo, la relativa a salud y salud mental. También vimos investigación que es manifiestamente cuestionable como la relativa al perfilado de usuarios en términos de orientación sexual, etnicidad, punto de vista religioso y político, etc.
- **¿Cómo hemos llegado a esta situación?** Fundamentalmente porque los investigadores de todo el mundo intentan estar a la par con los de EE.UU. y aquellos (solo) deben conseguir aprobación ética de un IRB para sus proyectos (desde 1975) o, mejor aún, que dicho IRB declare que su investigación está exenta de dicha evaluación.
- **¿Cómo es posible que necesitar evaluación ética lleve a investigación poco ética?** Porque la práctica totalidad de investigación con trazas digitales se ha venido realizando bajo el amparo de la siguiente exención:

  "Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, ***if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified****, directly or through identifiers linked to the subjects."*

- Bajo esa premisa prácticamente cualquier contenido que un usuario haya colgado en la Web para su difusión pública se considera lícito para su uso en investigación (sin solitar su permiso, por supuesto).

Ética en la minería de trazas digitales

Investigación en la industria

- Ni que decir tiene, los estándares de la investigación en las empresas privadas no eran (¿son?) mucho mejores.
- Por ejemplo, Facebook realizó en 2013 un experimento para manipular las emociones de casi 700.000 usuarios.
- La revista que publicó esa investigación emitió después una "expresión de preocupación" sobre la misma y otros investigadores criticaron la ética del experimento [1][2][3] Post inmenso con información sobre este experimento.
- Curiosamente, en 2012 Facebook había publicado los resultados de un experimento para incitar al voto a 61 millones de usuarios en las elecciones al Congreso de EE.UU. de 2010 sin que hubiese mucho revuelo... 🤷
- ¿Qué amparo ético tuvo el experimento del contagio de emociones? Según Facebook, que los usuarios hubiesen aceptado los términos de uso de la plataforma significaba que habían dado su consentimiento. Según la Universidad de Cornell, puesto que su personal no había tenido acceso directo a los datos el proyecto no necesitaba revisión del IRB. 🤦

Ética en la minería de trazas digitales

Investigación en la industria

- Otro escándalo surgió en 2014 cuando OKCupid (un sitio de citas) publicó con orgullo que hacía experimentos con sus usuarios, entre ellos:
  - El 15 de enero de 2013 eliminaron las fotos del sitio y comprobaron que, aunque hubo menos actividad en el sitio, aquellos usuarios que sí interactuaron lo hicieron con más intensidad y profundidad. A las 16:00 el sitio restauró las fotos y las conversaciones que se habían iniciado a ciegas terminaron abruptamente.
  - En otra ocasión manipularon el porcentaje de coincidencia para hacer que parejas que eran una "mala" opción (30% de coincidencia o menos) tratasen de concertar una cita (diciéndoles que tenían un 90% de coincidencia). Los usuarios manipulados iniciaron más conversaciones y, además, las mantuvieron en el tiempo. Puesto que esto podría implicar que su algoritmo era inútil probaron a sugerir que "buenas" coincidencias eran "malas". Aunque el número de conversaciones fue algo inferior siguió siendo superior a las conversaciones entre personas que tenían una baja coincidencia y, además, se les mostraba.
- Para mayor escarnio, en una entrevista posterior a esa publicación el CEO de OKCupid afirmó, al ser cuestionado sobre la supervisión ética del experimento, que ellos tenían "la farsa del consentimiento con los términos y condiciones".

Odds of a single message turning into a conversation

ACTUAL compatibility of users	number DISPLAYED to them: 30% match	60% match	90% match
30% match	10%	16%	17%
60% match	13%	13%	16%
90% match	16%	17%	20%

Ética en la minería de trazas digitales

Sobre los *Terms of Service* y el consentimiento informado

- Aunque se polemizó sobre el enfoque de Facebook (u OKCupid) al equiparar consentimiento informado con aceptar los términos de uso lo cierto es que la mayor parte de investigación académica posterior aceptó alegremente esa premisa [1]
- Lo cual es problemático por muchos motivos:
  - La mayor parte de usuarios (74%) se saltan la lectura de los términos de uso y cláuslas de privacidad [2].
  - Los que las "leen" lo hacen en alrededor de un minuto [2].
  - Aunque se lean, los documentos son básicamente incomprensibles [3].
  - Los usuarios han llegado a "aceptar" (al unirse a una plataforma ficticia) que sus datos se compartieran con la NSA o dar su primer descendiente como pago por el uso del servicio [2].
  - Una parte sustancial de los usuarios se han sentido coaccionados al aceptar los términos de uso ya que, de no hacerlo, no podrían acceder el servicio [4]
- Por otro lado, son minoría los usuarios que saben que sus contenidos pueden ser recolectados por investigadores [5] y la mayoría creen que no se debería poder hacer eso sin consentimiento previo [5][6] y sólo podrían publicarse de forma anónima [6] (Artículo interesante sobre cómo "disfrazar" datos obtenidos online para su publicación).
- Así, a día de hoy nadie afirma que aceptar los términos del servicio equivalga a consentimiento informado pero lo cierto es que tampoco sabemos cómo adaptar el consentimiento informado a los estudios con trazas digitales [7] y algunas de las recomendaciones que se hacen [8] son poco prácticas y, en muchos casos, inaplicables.
- Por ejemplo, sería difícil conseguirlo explícitamente si se deseasen estudiar la opiniones vertidas durante un debate político y hay que tener en cuenta, además, que se ha comprobado que, en aquellos casos en que los investigadores advierten a los miembros de una comunidad de que van a grabar sus interacciones y contenidos acaban por ser expulsados [9].

Ética en la minería de trazas digitales

¡Los datos son gente!

- Existe, sin embargo, un problema mayor que la pretensión de aceptar o no los términos de servicio como consentimiento informado.
- Se ha tratado de disociar los datos de las personas que hay tras ellos para así afirmar que **no** se trabaja con personas (tarea delicada) sino con datos (tarea más sencilla y aséptica).
- Obviamente, eso es inaceptable porque los datos son gente.
- En la misma línea de Rebecca Lemov está el muy recomendable capítulo *"Bring Back the Bodies"* por Catherine D'Ignazio y Lauren Klein del libro *"Data Feminism"*. También es muy pertinente el artículo de Sarah Gilbert, Jessica Vitak y Katie Shilton.
- ⚠ D'Ignazio y Klein señalan que términos que hemos venido usando en esta unidad como **"sesgo"** o ahora **"ética"** enfatizan la supuesta objetividad de la minería de datos y culpan a usuarios o sistemas individuales; argumentan que deberían usarse otros como **"opresión"** y **"justicia"** para subrayar que el problema es estructural.
- Por su interés dedicaremos más adelante algo de tiempo al concepto **FATE** *(Fairness, Accountability, Transparency and Ethics)* baste por ahora señalar que para hacer minería de trazas digitales debemos ser conscientes de que trabajamos con personas y no con datos.

Ética en la minería de trazas digitales

Principios maestros para la investigación con personas

- Como ya dijimos, en los 1970s, se desarrollaron en EE.UU. algunos de los principios que deberían guiar la investigación con seres humanos.
- En 1979 se publicó *The Belmont Report* que tuvo una enorme influencia en los años posteriores y que aún supone una guía inicial.
- El informe establece **3 principios éticos fundamentales:**
  - **Respeto a las personas:** proteger la autonomía de todas las personas y tratarlas con cortesía y respeto permitiendo el consentimiento informado. Los investigadores deben ser sinceros y no engañar;
  - **Benevolencia:** aplicar la filosofía de "no dañar" mientras se maximizan los beneficios para el proyecto de investigación y se minimizan los riesgos para los sujetos de investigación; y
  - **Justicia:** garantizar que se aplican de manera justa e igualitaria procedimientos razonables, no explotadores y ponderados.
- Ejercicio interesante: reflexionar sobre qué principios se violaron (además de la falta de consentimiento informado) en alguna de la investigación descrita en la introducción o en los experimentos de Facebook y OKCupid.

Ética en la minería de trazas digitales

Principios maestros para la investigación con personas

- Los principios de la *American Anthropological Association* pueden ser también de interés:
  - No dañar.
  - Ser abierto y honesto sobre el trabajo (de investigación).
  - Obtener el consentimiento informado y los permisos necesarios.
  - Sopesar las obligaciones éticas hacia colaboradores y partes afectadas.
  - Hacer accesible los resultados (de la investigación).
  - Proteger y preservar los registros.
  - Mantener relaciones profesionales respetuosas y éticas.
- Más información sobre estos principios aquí y sobre cuestiones éticas relativas a la antropología aquí.

Ética en la minería de trazas digitales

Principios maestros para la investigación con personas en Internet

- Tanto la *American Psychological Association* como la *British Psychological Sociecty* han elaborado abundante documentación sobre la forma de llevar a cabo investigación en Internet (aquí y aquí).
- La APA aporta un diagrama interesante para determinar si se precisa o no consentimiento informado [1]
- Por lo que respecta a la BPS sus principios podrían resumirse en los siguientes:
  - Respeto a la autonomía, privacidad y dignidad de las personas y comunidades.
  - Integridad científica.
  - Responsabilidad social.
  - Maximización de los beneficios y minimización del daño.

Ética en la minería de trazas digitales

Principios maestros para la investigación usando herramientas informáticas

- En cuanto que profesionales de la informática también sería de aplicación el código ético de la ACM:
  - Contribuir a la sociedad y al bienestar humano, reconociendo que todas las personas son partes interesadas en la informática.
  - Evitar dañar.
  - Ser honesto y digno de confianza.
  - Ser justo y no discriminar.
  - Respetar el trabajo necesario para producir nuevas ideas, inventos, trabajos creativos y artefactos informáticos.
  - Respetar la privacidad.
  - Honrar la confidencialidad.

Ética en la minería de trazas digitales

Principios maestros para la investigación en y sobre Internet

- Finalmente, la AoIR lleva casi dos décadas trabajando en los aspectos éticos relativos a la investigación **en** y **sobre** Internet.
- En este sentido produjo informes en 2002, 2012 y hace poco se aprobó la versión 3.0.
- Como futuros desarrolladores os pueden resultar, además, de interés leer de esa versión 3.0 los siguientes anexos:
  - *AI and Machine Learning: Internet Research Ethics Guidelines*.
  - *An "Impact Model" for ethical assessment*.
- Un primer punto de partida, antes de leer los informes completos, sería este diagrama.

Ética en la minería de trazas digitales

Principios maestros para la investigación en y sobre Internet

- Por lo que respecta a los principios rectores de la AoIR podrían resumirse del siguiente modo:
  - Cuanto más **vulnerable** es la comunidad (o participantes) mayores son las obligaciones del investigador.
  - **El daño se define contextualmente**, es decir, los principios éticos dependen del contexto.
  - **Los datos provienen de personas**. Así pues, si se trabaja con datos generados por personas se trabaja con personas.
  - Hay que **equilibrar** los **derechos** de las personas con los **beneficios** de la investigación.
  - **Los problemas éticos surgen a cada paso**: desde la planificación hasta la propia investigación y la difusión de los resultados (o de los datos recolectados).
  - **La toma de decisiones éticas es un proceso deliberativo** y es mejor consultar con diferentes personas y fuentes: colegas, personas que usan o están familiarizadas con los sitios que se estudian, los comités de revisión de investigación, las guías éticas (p.ej., la propia de la AoIR, o las ya mencionadas de APA, BPS o AAA), los estudios publicados y, cuando corresponda, los precedentes legales

Problemas ideológicos

- El análisis de datos masivos no solo es un campo científico-tecnológico, exhibe rasgos que pueden considerarse ideológicos, incluso mitológicos: el **dataísmo**.
- Algunos ejemplos:
  - *"inanimate data can never speak for themselves, and we always bring to bear some conceptual framework, either intuitive and ill-formed, or tightly and formally structured, to the task of investigation, analysis, and interpretation."*
  - *"the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy"*
  - *"The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning. Like Caesar, we may construe them in self-serving ways that are detached from their objective reality."*
  - *"Any statistical test or machine learning algorithm expresses a view of what a pattern or regularity is and any data has been collected for a reason based on what is considered appropriate to measure. One algorithm will find one kind of pattern and another will find something else. One data set will evidence some patterns and not others. Selecting an appropriate test depends on what you are looking for."*
  - *"Dataism thrives on the assumption that gathering data happens outside any preset framework—as if Twitter facilitates microblogging just for the sake of generating "life" data—and data analysis happens without a preset purpose—as if data miners analyze those data just for the sake of accumulating knowledge about people's behavior."*
  - *"Dataists believe that humans can no longer cope with the immense flows of data, hence they cannot distil data into information, let alone into knowledge or wisdom. The work of processing data should therefore be entrusted to electronic algorithms, whose capacity far exceeds that of the human brain. In practice, this means that Dataists are sceptical about human knowledge and wisdom, and prefer to put their trust in Big Data and computer algorithms."*

Problemas ideológicos

- Tan importante como reconocer las limitaciones de los datos con los que se trabaja y las consideraciones éticas o de justicia que merecen las personas que los producen es no caer en una perspectiva "dataísta" según la cual los datos y su análisis permite un acceso objetivo y neutro a La Verdad puesto que en cualquier decisión hay un sustrato ideológico.
- Así, algo en apariencia trivial como crear una lista de palabras "ofensivas" [1][2] tiene una base ideológica: qué resulta sucio u obsceno y qué no y, en consecuencia, qué tipo de textos (y comunidades) van a ser invisibilizados e ignorados.
- Por último, hay que tener en cuenta que un enfoque **basado en datos** corre el riesgo de convertirse en un enfoque **dirigido por los datos** y, si además se aplica con ánimo exploratorio cualquier método de análisis disponible con el objetivo de descubrir "algo" (lo que sea), podemos acabar con un enfoque de *Data Piñata* que es, obviamente, una aproximación acientífica.

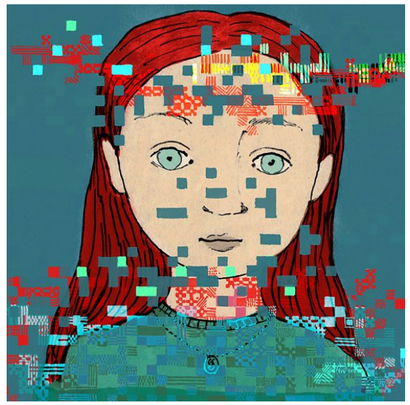
Fairness, Accountability, Transparency and Ethics

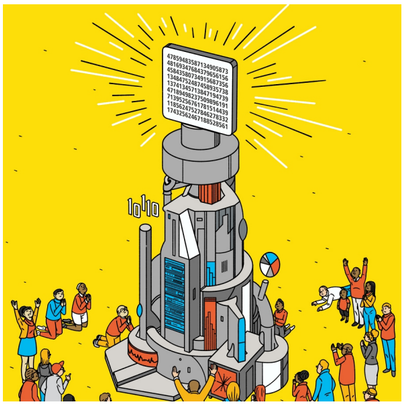
- Durante los últimos años ha aumentado la preocupación por los aspectos no solo éticos sino de justicia y no discriminación en la aplicación de técnicas de aprendizaje automático y minería de datos.
- Así, entre 2014 y 2018 se celebró el congreso FAT/ML y desde 2018 hasta la fecha el congreso FAT*.
- El objetivo de dichos congresos es la investigación multidisplinar en sistemas socio-técnicos con énfasis en cuestiones de justicia, responsabilidad y transparencia.
- En marzo de 2017 Microsoft estableció el grupo FATE *(Fairness, Accountability, Transparency, and Ethics in AI)* y parece que ese acrónimo ha gustado a la comunidad.
- Las perspectivas FATE tratan de analizar las consecuencias de los sistemas inteligentes y la ciencia de datos sobre la sociedad y los individuos para desarrollar tecnologías basadas en dichos sistemas que sean justas y no discriminatorias.
- Los siguientes recursos pueden resultar de interés para iniciarse en esta perspectiva:
  - El libro (aún incompleto) *Fairness and machine learning*
  - Los materiales del grupo *Fairness, accountability, transparency and ethics (FATE) Reading & Debate Group* de la UPF.
  - El tutorial de Sara Hajian, Francesco Bonchi y Carlos Castillo en KDD'16 (incluye vídeos).
  - La colección de ensayos *"Data and discrimination: collected essays"* del *Open Technology Institute*.
  - *The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making*.
  - *If AI is the problem, is debiasing the solution?*.
- ⚠ **Las empresas tecnológicas intentarán co-optar la ética (véase aquí o aquí) por lo que es crucial reforzar el concepto de justicia.**

Para saber más...

- La tesis doctoral de Alexandra Olteanu y su artículo con Castillo, Diaz y Kiciman.
- Los artículos:
  - *A Total Error Framework for Digital Traces of Humans*
  - *Unpacking the Expressed Consequences of AI Research in Broader Impact Statements*
  - *Stewardship of global collective behavior*.
- Los siguientes libros:
  - *"Bit By Bit: Social Research in the Digital Age"*
  - *"Data feminism"*
  - *"Ethics of Big Data: Balancing Risk and Innovation"*
  - *"Fairness and Machine Learning: Limitations and Opportunities"*
  - *"Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy"*
  - *"The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences"*
- La entrada *Internet Research Ethics* en la *Stanford Encyclopedia of Philosophy*.

Para saber más...





Conclusiones

- La minería de trazas digitales prometía una oportunidad única para entender mejor a los individuos y a la sociedad en su conjunto.
- Existía la creencia de que la explotación de que cantidades ingentes de datos generados *online* bastaría para un acceso objetivo a la realidad.
- Esa creencia (dataísmo) ignora las múltiples limitaciones de los datos debidas a sesgos demográficos, de autoselección, en la producción de los contenidos y los propios del mundo *offline* (sexismo, racismo, homofobia, etc).
- Esos sesgos y actitudes discriminatorias acaban por ser incorporadas en los sistemas desarrollados que las reproducen y refuerzan.
- El dataísmo también ignora que los datos se generan en un entorno adversarial que es manipulado por múltiples agentes con intenciones espurias.
- Por otro lado, el análisis de trazas digitales plantea múltiples retos éticos que no pueden solventarse mediante los términos de uso del servicio.
- Por último, un simple barniz ético es insuficiente y se hace necesaria una perspectiva más completa que persiga la justicia y no discriminación.