Transcripcions en espanyol i català.

El futuro de la visualización de datos

«TRAZANDO UN CAMINO HACIA EL FUTURO. EL FUTURO DE LA VISUALIZACIÓN DE DATOS. JEFFREY HEER. TRIFACTA. UNIVERSIDAD DE WASHINGTON»

«STRATA+HADOOP. JEFFREY HEER, CO-FUNDADOR Y PROFESOR. TRIFACTA. UNIVERSIDAD DE WASHINGTON»

En las conversaciones habituales, las discusiones sobre visualizaciones de datos suelen referirse a ejemplos como este: gráficos diseñados a medida que con frecuencia pretenden transmitir una historia y son el trabajo manual de diseñadores de visualizaciones de gran talento.
A lo largo de los años que he dedicado a investigar la visualización, mis estudiantes y yo hemos trabajado para diseñar herramientas que permitan crear gráficos tan sofisticados como estos, desarrollando herramientas que se han hecho populares, como Protovis, Vega y D3.
Y aunque nos alegra el éxito que ha acompañado a estas herramientas, creo que son solo una pequeña parte del amplio ecosistema de las visualizaciones. Si lo pensamos, la mayoría de las visualizaciones no son el resultado de un código escrito a mano, sino fruto de herramientas para el usuario final. Y con frecuencia, vemos visualizaciones que tienen este aspecto... o aplicaciones como estas.
Y aunque su intención sea buena, muchas de las herramientas para los usuarios no tienen en cuenta los principios de la percepción o fracasan a la hora de facilitar el proceso de análisis y exploración visual.
Por eso, cuando pienso en el futuro de la visualización de datos y en cómo trazar ese camino, una de las cosas a tener en cuenta es cómo pasar de las herramientas que funcionan para los diseñadores a las herramientas que permiten el análisis para los que toman las decisiones, y que estos defiendan sus causas para una mejor toma de decisiones entre la industria y el gobierno, por ejemplo.
¿Cómo podemos mejorar la tecnología más reciente? Creo que una de las formas es incorporar más opciones de diseño en nuestras herramientas de visualización.
Para que entendáis a qué me refiero, vamos a realizar un experimento. Os voy a enseñar unas formas y quiero que las comparéis. No gritéis las respuestas, luego haremos una rápida encuesta. Aquí hay dos círculos y quiero que comparéis su área. ¿Cuántas veces es el círculo grande mayor que el pequeño? Levantad la mano si creéis que es cuatro veces más grande. Vale, ¿Cinco? ¿Seis? ¿Siete? ¿Ocho? ¿Nueve, diez...? Mucha gente. ¿Once o más? Algunas personas.

Vale. Veamos otro ejemplo. Comparad el tamaño de estas barras. ¿Cuánto mayor es la barra más grande?
Levantad la mano si creéis que es cuatro veces más grande. Nadie. ¿Cinco, seis? Algunos más. ¿Siete? Aún más. ¿Ocho? Mucha gente. ¿Nueve? Muchos menos. ¿Diez? ¿Once o más? Casi nadie.
La respuesta es la misma en ambos casos. Es siete veces mayor. Pero al mirar las respuestas lo que vemos es que a pesar de que la diferencia entre las áreas es la misma, había más precisión y menos variabilidad al comparar la longitud que el área. Aunque esta sea una encuesta poco científica, mi equipo y otros han llevado a cabo experimentos similares para comparar la precisión de distintos tipos de codificación. Aquí podéis ver las tasas de error al comparar cosas como la posición, la longitud, angulación y el área.

«PRECISIÓN EN LA DECODIFICACIÓN VISUAL»

Y combinando los resultados de estos experimentos, podemos construir *rankings* sobre la efectividad visual de distintas codificaciones en distintos tipos de tareas de percepción. Por ejemplo, cuando se comparan cantidades, la posición y la longitud funcionan mejor que el ángulo o el área, que funcionan solo un poco mejor que la codificación por colores.
Esto es útil como guía, no como reglamento, de lo que hacen los diseñadores. Pero también es útil para orientar los algoritmos, que automáticamente podrían recomendar el gráfico más útil y eficaz.
Por ejemplo, en Trifacta nuestros productos incluyen soporte técnico para el diseño visual de grandes bases de datos. Aquí vemos la visualización de un mapa que muestra la procedencia de las contribuciones a los partidos políticos. En este caso, utiliza una usual codificación basada en el color para los estados. ¿Os dais cuenta del problema? Como acabo de mencionar, el color es uno de los canales de codificación menos precisos a la hora de comparar datos cuantitativos. Pero en realidad hay un problema aún mayor con este gráfico, y es que nuestra percepción está relacionada con la forma y el tamaño de los estados. Por lo tanto, en realidad no vemos lo que sucede en Washington DC, que resulta ser la fuente de muchas contribuciones.
¿Qué otras codificaciones podríamos barajar? Por ejemplo, queremos mantener el contexto espacial del mapa, así que la posición ya está determinada, pero podríamos introducir algo más preciso, como el área, para permitir comparaciones más eficaces. Pero, como he dicho, la posición es una de las codificaciones más efectivas, así que vamos a mejorar esta disposición con gráficos de barras a la izquierda, para así poder hacer comparaciones y conectar estas visiones mediante la interacción. Eso nos permite más exploración, poder ver más patrones y hacer comparaciones más precisas.
Así que incorporar mejores opciones de diseño en nuestras herramientas es uno de los caminos, pero es solo el principio. Creo que es hora de que nos replanteemos algunos de los interfaces básicos para las visualizaciones de datos. Por ejemplo, la mayoría de las herramientas incluyen un proceso para la especificación gráfica. ¿Cómo podemos convertirlo en un proceso de rápida exploración?

TRAZANDO EL CAMINO HACIA EL FUTURO
DE DISEÑADORES A RESPONSABLES DE LA TOMA DE DECISIONES
DE LA ESPECIFICACIÓN A LA EXPLORACIÓN

Las herramientas para usuario más comunes permiten seleccionar los subconjuntos de datos que nos interesan y luego elegir entre una selección de gráficas o codificaciones visuales para construir una representación de forma manual. Pero esto puede ser extremadamente tedioso y también parece una oportunidad desaprovechada de recomendar vistas interesantes.
Por ejemplo, en el Trifacta Video Profiler... Ahí está. No hace falta que la gente cree los gráficos personalmente, nosotros los suministramos. A la izquierda podéis ver esquemas donde presentamos resúmenes de todas las dimensiones presentes en una base de datos. Y el usuario puede abrir la selección para ver más detalle. Por ejemplo, este panel de la derecha nos muestra cuándo se dan las contribuciones a partidos políticos. Se muestra también en una línea cronológica, así como en gráficas resumidas por periodos, tales como el día de la semana, el día del mes, etc.
Y al hacer esto vemos una serie de patrones: vemos que las contribuciones se dan con regularidad y en aumento a medida que se acercan las elecciones, y que son más frecuentes en días laborables o al final del mes. Así podemos observar algo útil de forma inmediata, sin tener que especificar los gráficos.
Pero lo más importante es que estas representaciones se han ofrecido automáticamente en función de los datos. Por ejemplo, hay muy poca variación en lo que respecta a horas, minutos y segundos, así que no mostramos esas gráficas para evitar distraer al usuario o hacerle perder el tiempo.

De forma paralela, en mi laboratorio de datos interactivo, en la universidad de Washington, estamos explorando nuevas herramientas de exploración para el usuario final. Por ejemplo, nuestro sistema Data Voyager hace búsquedas entre miles de visualizaciones y las ordena, según valoraciones estadísticas y perceptuales para recomendar visualizaciones para conjuntos de datos.
Y el usuario también está incluido en el proceso. Por ejemplo, puedes dirigir las recomendaciones y actualizar la muestra indicando tus campos de interés. Así que estamos cambiando la manera en la que interactuamos y exploramos los datos para permitir una exploración más abierta y más rápida.

DE DISEÑADORES A RESPONSABLES DE LA TOMA DE DECISIONES
DE LA ESPECIFICACIÓN A LA EXPLORACIÓN

Hay muchos retos que surgen en este cambio de la especificación a la exploración que no estoy mencionando, porque para hacerlo bien necesitaría los talentos combinados de la comunidad de Strata: del diseño de visualizaciones, de estadísticas y aprendizaje automatizado, así como de los sistemas de *big data*. Y estoy muy emocionado ante estos desafíos, pero también hay que considerar cómo estas herramientas mejorarán el análisis de la forma más productiva. Y para ilustrar este punto quiero compartir con

vosotros un conjunto de datos que suelo dar a mis estudiantes de visualización. Es un conjunto de datos clásico, publicado por primera vez en los años 50, que compara el efecto de los antibióticos –que en aquel momento eran los fármacos estrella– mostrando su efecto sobre una serie de cepas bacterianas.
Les doy este conjunto de datos a los estudiantes y les pido que exploren los datos, que creen un gráfico que dé respuesta a alguna pregunta interesante.
Y estos son algunos de los trabajos de los estudiantes. Podéis comprobar la gran variedad de diseños que han explorado. Es muy interesante, pero algo llamativo que surge cuando empiezas a analizar estos gráficos de cerca es que aunque hay mucha variación en los diseños, todos están respondiendo a la misma pregunta: "¿Qué antibiótico deberíamos usar?". Y la fijación en esa pregunta me recuerda a una de mis máximas preferidas del experto en visualización Eduard Tufte, quien nos aconseja mostrar "variación en los datos, no solo variación en el diseño".
La idea es que aunque puede ser útil mostrar múltiples representaciones del mismo subconjunto de datos, a menudo es más interesante explorar diferentes secciones y transformaciones de los datos que provoquen nuevas preguntas o sugieran nuevas hipótesis. O, dicho de otra forma, deberíamos pensar en cómo hacer que nuestras herramientas estimulen el ejercicio del escepticismo sobre los datos y ayuden a considerar nuevas preguntas.
Así que vamos a ver esta visualización alternativa. Es bastante sencilla. Lo que vemos son diferentes bacterias en un diagrama de dispersión. Las bacterias están coloreadas según su género y están distribuidas de forma acorde a la efectividad de dos antibióticos: neomicina y penicilina. La esquina inferior izquierda muestra las áreas de fuerte resistencia y en la esquina superior derecha la baja resistencia.
Y la idea es que la mera sugestión de esta visualización relativamente sencilla podría movernos a considerar otras preguntas. Podríamos preguntarnos "¿Qué nos dice la respuesta a los antibióticos sobre la biología de las bacterias?". Le estamos dando la vuelta a la pregunta.
Volviendo a la gráfica, quizá hayáis percibido algo interesante y es que hay diferentes grupos de bacterias, pero los géneros están distribuidos de manera un tanto sorprendente. ¿No sería de esperar que las bacterias de la misma familia estuvieran más agrupadas? Si esto os hace dudar de los datos, tenéis toda la razón, ya que de hecho los datos contienen errores. La comunidad científica se equivocó y hay dos clasificaciones erróneas en este conjunto de datos.
Tuvieron que pasar varias décadas tras la publicación de estos datos para que la comunidad científica detectara estos errores. Y sin embargo las pruebas estuvieron ante nuestras narices todo el tiempo, si hubiéramos presentado los datos bajo este prisma en particular.
Así que es interesante pensar en cómo las herramientas pueden llevarnos a considerar los datos de manera más amplia. Y en nuestra evolución de la especificación a la exploración deberíamos tener en cuenta de qué manera este cambio va a mejorar el análisis. Por ejemplo, priorizando la variación en los datos sobre la variación en el diseño.
En conclusión, me gustaría augurar un futuro en el que nuestras herramientas no solo ayuden a construir las visualizaciones, sino que ayuden a explorar los datos de forma

más rica y, quizá, en última instancia conduzcan a mejores percepciones y mejores decisiones. Y espero que todos nosotros construyamos juntos ese futuro. Gracias.

El futur de la visualització de dades

«TRAÇANT UN CAMÍ CAP AL FUTUR. EL FUTUR DE LA VISUALITZACIÓ DE DADES. JEFFREY HEER. TRIFACTA. UNIVERSITAT DE WASHINGTON»

«STRATA+HADOOP. JEFFREY HEER, COFUNDADOR I PROFESSOR. TRIFACTA. UNIVERSITAT DE WASHINGTON»

En les converses habituals, les discussions sobre visualització de dades solen referir-se a exemples com aquest: gràfics dissenyats a mida que sovint pretenen transmetre una història i són la feina manual de dissenyadors de visualitzacions amb molt de talent.
Al llarg dels anys que he dedicat a investigar la visualització, els meus estudiants i jo hem treballat per dissenyar eines que permetin crear gràfics tan sofisticats com aquests i hem desenvolupat eines que s'han fet populars com Protovis, Vega i D3.
I tot i que ens alegra l'èxit que ha acompanyat aquestes eines, crec que només són una petita part de l'ampli ecosistema de les visualitzacions. Si ho pensem, la majoria de les visualitzacions no són el resultat d'un codi escrit a mà, sinó fruit d'eines per a l'usuari final. I sovint veiem visualitzacions que tenen un aspecte com aquest o aplicacions com aquestes.
I encara que la intenció és bona, moltes de les eines per als usuaris no tenen en compte els principis de la percepció o fracassen a l'hora de facilitar el procés d'anàlisi i exploració visual.
Per això, quan penso en el futur de la visualització de dades i en com traçar aquest camí, una de les coses que cal tenir en compte és com passar de les eines que funcionen pels dissenyadors a les eines que permeten l'anàlisi per aquells que prenen les decisions. I que aquests defensin les seves causes per a una millor presa de decisions entre la indústria i el govern, per exemple.
Com podem millorar la tecnologia més recent? Crec que una de les maneres és incorporar més opcions de disseny a les nostres eines de visualització. Perquè entengueu a què em refereixo, farem un experiment ràpid. Us ensenyaré unes formes i vull que les compareu. No em digueu la resposta, després farem una enquesta ràpida.
Aquí hi ha dos cercles i vull que en compareu l'àrea. Quantes vegades el cercle gran és més gran que el petit? Aixequeu la mà si creieu que és quatre vegades més gran.
D'acord. Cinc? Sis? Set? Vuit? Nou, deu...? Molta gent. Onze o més? Algunes persones.
D'acord. Vegem un altre exemple. Compareu la mida d'aquestes barres. Quantes vegades més gran és la barra més gran?
Aixequeu la mà si creieu que és quatre vegades més gran. Ningú. Cinc, sis? Algú més. Set? Encara més. Vuit? Molta gent. Nou? Molts menys. Deu? Onze o més? Gairebé ningú.

La resposta és la mateixa en ambdós casos. És set vegades més gran. Però quan analitzem les vostres respostes el que veiem és que, tot i que la diferència de les àrees és la mateixa, hi havia més precisió i menys variabilitat quan es comparava la longitud que quan es comparava l'àrea. Encara que aquesta enquesta sigui poc científica, el meu equip i altres han realitzat experiments semblants per comparar la precisió de diferents tipus de codificació. Aquí podeu veure les taxes d'error a l'hora de comparar coses com la posició, la longitud, l'angulació i l'àrea.

«PRECISIÓ EN LA DESCODIFICACIÓ VISUAL»

I combinant els resultats d'aquests experiments, podem construir rànquings sobre l'efectivitat visual de diferents codificacions en diferents tipus de tasques de percepció. Per exemple, quan es comparen quantitats, la posició i la longitud funcionen millor que l'angle o l'àrea, que funcionen només una mica millor que la codificació per colors. Això és útil com a guia, no com a reglament, del que fan els dissenyadors. Però també és útil per orientar els algoritmes, que automàticament podrien recomanar el gràfic més útil i eficaç.
Per exemple, a Trifacta els nostres productes inclouen suport tècnic per al disseny visual de grans bases de dades. Aquí veiem la visualització d'un mapa que mostra la procedència de les contribucions a partits polítics. En aquest cas, utilitza una codificació usual basada en el color per estats. Us adoneu del problema? Com acabo de comentar, el color és un dels canals de codificació menys precisos a l'hora de comparar dades quantitatives. Però en realitat encara hi ha un problema més gran amb aquest gràfic. I és que la nostra percepció està relacionada amb la forma i la mida dels estats. Per tant, en realitat no veiem què passa a Washington DC, que resulta ser la font de moltes contribucions.
Quines altres codificacions podríem fer servir? Per exemple, volem mantenir el context espacial del mapa, així que la posició ja està determinada, però podríem introduir alguna cosa més precisa com l'àrea per permetre comparacions més eficaces. Com he dit, però, la posició és una de les codificacions més efectives, així que millorarem aquesta disposició amb gràfics de barres a l'esquerra per tal de poder fer comparacions i connectar aquestes visions mitjançant la interacció. Això ens permet més exploració, poder veure més patrons i fer comparacions més precises.
Així que incorporar millors opcions de disseny a les nostres eines és un dels camins, però només és el principi. Crec que és hora de replantejar-nos alguna de les interfícies bàsiques per a les visualitzacions de dades. Per exemple, la majoria de les eines inclouen un procés per a l'especificació gràfica. Com el podem convertir en un procés d'exploració ràpida?

TRAÇANT EL CAMÍ CAP AL FUTUR
DE DISSENYADORS A RESPONSABLES DE LA PRESA DE DECISIONS
DE L'ESPECIFICACIÓ A L'EXPLORACIÓ

Les eines per a l'usuari més comunes permeten seleccionar els subconjunts de dades que ens interessen i després elegir entre una selecció de gràfiques o codificacions

visuals per construir una representació de forma manual. Però això pot ser extremadament tediós i també sembla una oportunitat desaprofitada per recomanar vistes interessants.
Per exemple, al Trifacta Video Profiler... Aquí el tenim. No cal que la gent creï els gràfics personalment, nosaltres els els subministrem. A l'esquerra podeu veure esquemes on presentem resums de totes les dimensions presents en una base de dades. I l'usuari pot obrir la selecció per veure-ho amb més detall. Per exemple, a la dreta veiem quan es fan les contribucions a partits polítics. Aquesta informació també es mostra en una línia cronològica, així com en gràfiques resumides per períodes tals com el dia de la setmana, el dia del mes, etc. I quan fem això veiem una sèrie de patrons. Veiem que les contribucions es donen amb regularitat i augmenten a mesura que s'acosten les eleccions i que són més freqüents els dies laborables o al final del mes. Així podem observar alguna cosa útil de forma immediata, sense haver d'especificar els gràfics.
Però el més important és que aquestes representacions s'han ofert automàticament en funció de les dades. Per exemple, hi ha molt poca variació respecte a les hores, minuts i segons, així que no mostrem aquestes gràfiques per evitar distreure l'usuari o fer-li perdre el temps.
Paral·lelament, estem explorant noves eines d'exploració per a l'usuari final al meu laboratori de dades interactiu a la Universitat de Washington. Per exemple, el nostre sistema Data Voyager fa cerques entre milers de visualitzacions i les ordena segons valoracions estadístiques i perceptives per recomanar visualitzacions per conjunts de dades.
I l'usuari també està inclòs en el procés. Per exemple, pots dirigir les recomanacions i actualitzar la mostra indicant els teus camps d'interès. Així que estem canviant la manera amb què interactuem i explorem les dades per permetre una exploració més oberta i més ràpida.

DE DISSENYADORS A RESPONSABLES DE LA PRESA DE DECISIONS
DE L'ESPECIFICACIÓ A L'EXPLORACIÓ

Sorgeixen molts reptes en aquest canvi de l'especificació a l'exploració que no he mencionat, perquè per fer-ho bé necessitaria els talents combinats de la comunitat de Strata: del disseny de visualitzacions, d'estadístiques i aprenentatge automatitzat, així com dels sistemes de dades massives. I estic molt emocionat davant d'aquests desafiaments, però també cal considerar com aquestes eines milloraran l'anàlisi d'una manera més productiva. I per il·lustrar-ho, vull compartir amb vosaltres un conjunt de dades que solo donar als meus estudiants de visualització. És un conjunt de dades clàssic, publicat per primera vegada els anys 50, que compara l'efecte dels antibiòtics –que en aquell moment eren els fàrmacs estrella– i mostra el seu efecte sobre una sèrie de soques bacterianes.
Dono aquest conjunt de dades als estudiants i els demano que explorin les dades, que creïn un gràfic que doni resposta a alguna pregunta interessant.
I aquests són alguns dels treballs dels estudiants. Podeu comprovar la gran varietat de dissenys que han explorat. És molt interessant, però una cosa que crida l'atenció quan

comences a analitzar aquests gràfics de prop és que, tot i que hi ha molta variació en els dissenys, tots responen a la mateixa pregunta: “Quin antibiòtic hauríem de fer servir?” I la fixació amb aquesta pregunta em recorda una de les meves màximes preferides de l’expert en visualització Eduard Tufte, que ens aconsella mostrar “variació en les dades, no només variació en el disseny.”

La idea és que, encara que pot ser útil mostrar múltiples representacions del mateix subconjunt de dades, sovint és més interessant explorar diferents seccions i transformacions de les dades que provoquin preguntes noves o suggereixin hipòtesis noves. O dit d’una altra manera, hauríem de pensar en com fer que les nostres eines estimulin l’exercici de l’escepticisme sobre les dades i ajudin a considerar preguntes noves.

Així que vegem aquesta visualització alternativa. És força senzilla. El que veiem són diferents bactèries en un diagrama de dispersió. Les bactèries estan acolorides segons el gènere i estan distribuïdes d’acord amb l’efectivitat de dos antibiòtics: neomicina i penicil·lina. La cantonada inferior esquerra mostra les àrees de forta resistència i la cantonada superior dreta, les de baixa resistència.

I la idea és que la mera suggestió d’aquesta visualització relativament senzilla podria fer-nos considerar altres preguntes. Ens podríem preguntar: “Què ens revela la resposta als antibiòtics sobre la biologia de les bactèries?” Estem girant la pregunta. Tornant a la gràfica, potser heu percebut una cosa interessant i és que hi ha diferents grups de bactèries, però els gèneres estan distribuïts de manera una mica sorprenent. No semblaria que les bactèries de la mateixa família haurien d’estar més agrupades? Si això us fa dubtar de les dades, teniu tota la raó. De fet, les dades contenen errors. La comunitat científica es va equivocar i hi ha dues classificacions errònies en aquest conjunt de dades.

Van haver de passar diverses dècades després de la publicació d’aquestes dades perquè la comunitat científica detectés aquests errors. I això que vam tenir la prova davant els nassos tot aquest temps, si haguéssim presentat les dades sota aquest prisma en particular.

Així que és interessant pensar en com les eines poden portar-nos a considerar les dades d’una manera més àmplia. I en la nostra evolució de l’especificació a l’exploració hauríem de tenir en compte de quina manera aquest canvi millorarà l’anàlisi. Per exemple, prioritzant la variació en les dades sobre la variació en el disseny.

En conclusió, m’agradaria augurar un futur en què les nostres eines no només ajudin a construir les visualitzacions, sinó que ajudin a explorar les dades de manera més rica i potser, en última instància, condueixin a millors percepcions i millors decisions. I espero que tots nosaltres construïm junts aquest futur. Gràcies.