MÈTODE Science Studies Journal (2014). Universitat de València.

DADES MASSIVES I ESTADÍSTICA

LA PERSPECTIVA D'UN ESTADÍSTIC

DAVID ROSSELL

Les dades massives (*big data*) representen un recurs sense precedents per a afrontar reptes científics, econòmics i socials, però també incrementen la possibilitat de traure conclusions enganyoses. Per exemple, l'ús d'enfocaments basats exclusivament en dades i que es despreocupen de comprendre el fenomen en estudi, que s'orienten a un objectiu esmunyedís i canviant, que no tenen en compte problemes determinants en la recopilació de dades, que resumeixen o «cuinen» inadequadament les dades i que confonen el soroll amb el senyal. Repassarem alguns casos reeixits i il·lustrarem com poden ajudar els principis de l'estadística a obtenir una informació més fiable de les dades. També abordarem els reptes actuals que requereixen estudis metodològics dinàmics, com les estratègies d'eficiència computacional, la integració de dades heterogènies, estendre els fonaments teòrics a qüestions cada vegada més complexes i, potser el més important, formar una nova generació de científics capaços de desenvolupar i implantar aquestes estratègies.

Paraules clau: dades massives, estadística, estudis de cas, trampes, reptes.

QUÈ SÓN LES DADES MASSIVES?

En els últims anys s'ha produït un increment significatiu en la nostra capacitat d'aplegar, emmagatzemar i compartir dades. Segons IBM, el 90 % de les dades del món s'ha generat en els últims dos anys (International Business Machines Corporation, 2011). Aquestes dades procedeixen d'internet (cerques, xarxes socials, blogs, imatges), telèfons d'última generació, estudis científics (genòmica, imatges cerebrals, epidemiologia, medi ambient), negocis (dades de clients, transaccions, indicadors financers), administració (població, salut, clima, sensors automàtics) i altres fonts.

La importància estratègica de les dades massives no consisteix en la quantitat sinó en les aplicacions potencials que ofereixen. Vegem alguns exemples. La caracterització de malalties complexes a escala molecular combinada amb l'historial mèdic i de tractament i amb proves diagnòstiques o d'imatge ofereix oportunitats sense precedents per a personalitzar la medicina. El Gran Col·lisionador d'Hadrons registra dades 40 milions de vegades per segon per a comprovar les teories de la física. Els llocs web generen cada dia milions de recomanacions, de comparances entre nous productes i entre els seus preus. Les dades poden ajudar a gestionar ciutats o recursos naturals, a estudiar el canvi climàtic o a promoure el desenvolupament de regions. Les notes en blogs i xarxes socials s'aprofiten per a dissenyar estratègies polítiques i per a estudiar com es difonen les idees.

> «L'EXPERIÈNCIA HA ENSENYAT ALS ESTADÍSTICS QUE LES DADES PODEN SER ENGANYOSES I, PITJOR ENCARA, DONAR UNA SENSACIÓ ERRÒNIA D'OBJECTIVITAT»

Gràcies a l'ampli abast de tot aquest potencial, els mitjans de comunicació, el món acadèmic i el dels negocis han acollit les dades massives amb un entusiasme fronterer a vegades amb el sensacionalisme. Termes com ara *allau de dades* o *tsunami* s'han fet comuns. El Fòrum Econòmic Mundial de 2012 va declarar les dades com un nou tipus d'actiu econòmic comparable a la moneda o a l'or (Fòrum Econòmic Mundial, 2012). Les professions relacionades amb la gestió de dades encapçalen constantment moltes classificacions. Deixant de banda el bombo publicitari, revisarem tant els èxits com les limitacions i destacarem les lliçons apreses i els reptes

pendents. Encara que les dades massives requereixen un enfocament pluridisciplinari, adoptarem un punt de vista estadístic. L'estadística és una disciplina dedicada específicament a recopilar, analitzar i interpretar dades. És a dir, ens porta de les preguntes a les dades, de les dades a la informació i de la informació al coneixement i a la presa de decisions. Pot semblar sorprenent, doncs, que els estadístics hagen estat relativament cautelosos a l'hora d'acollir les dades massives com una força totpoderosa. Al meu parer l'explicació és senzilla. L'experiència ha ensenyat als estadístics que les dades poden ser enganyoses i, pitjor encara, donar una sensació errònia d'objectivitat. Encara que siguen poderoses, les dades massives també obren la porta a moltes confusions. A causa de la varietat d'aplicacions (les dades massives sovint es defineixen com les tres V: volum, velocitat i varietat) no podrem abastar tot el que s'hi refereix, per això em limitaré a abordar alguns dels principals problemes i a posar alguns exemples.

Maximilien Brice (2009 CERN)

La importància estratègica de les dades no consisteix en la quantitat sinó en els usos potencials que ofereixen. Per exemple, el Gran Col·lisionador d'Hadrons registra dades 40 milions de vegades per segon per a comprovar les teories de la física.

LES DADES I EL PROCÉS SUBJACENT

L'aplicació que el mànager de beisbol Billy Beane va fer dels indicadors de rendiment i d'anàlisi de dades per a formar un equip competitiu (Lewis, 2003) s'ha convertit ja en tot un clàssic contemporani dels casos d'aprofitament reeixit de les dades que fins i tot va donar lloc a una pel·lícula de Hollywood bastant popular. El mèrit més notable de Bean és que el seu equip jugava millor que rivals amb major pressupost i dirigits per experts en beisbol. L'extraordinària precisió dels sondejos electorals britànics (Curtice i Firth, 2008) i nord-americans (Silver, 2012), que va triturar les previsions dels analistes polítics, és un altre èxit recent. Altres casos són els de les prediccions meteorològiques que anunciaven catàstrofes naturals (Silver, 2012), o l'explosió de les tecnologies *-òmiques* en les quals es basen molts, si no la majoria, dels avenços recents en biomedicina.

> «LES NOVES TECNOLOGIES SÓN INÚTILS A MENYS QUE CIENTÍFICS BRILLANTS PLANTEGEN PREGUNTES RELLEVANTS I INTERPRETEN ELS RESULTATS EN EL CONTEXT APROPIAT»

Aquestes històries potser han donat la falsa impressió que amb les dades ja ens està bé. Per exemple, en una entrevista publicada per *The New York Times* s'afirmava que les dades poden reemplaçar l'experiència i la intuïció, la qual cosa afavoreix un enfocament més científic (Lohr, 2012). No podria estar menys d'acord amb aquest punt de vista, que il·lustra un possible problema de les dades massives. Si bé és cert que les opinions no contrastades amb dades poden conduir a conclusions errònies, també les anàlisis cegues menen a error ben sovint. Disposar de dades fiables i de coneixements ferms, lluny d'oposar-se, es complementen. En els anteriors exemples, les prediccions van tenir èxit perquè estudiaven sistemes fonamentalment reproduïbles, i implicaven la comprensió del fenomen que estudiaven. Les variables triades per a predir el rendiment en el beisbol es prestaven a una interpretació natural de la matèria d'estudi. I els pronòstics de Silver aprofitaven els seus coneixements sobre la política nord-americana. Les prediccions meteorològiques es basen en simulacions informàtiques i lleis físiques, que els meteoròlegs corregeixen posteriorment per a eliminar les imprecisions sistemàtiques. Les noves tecnologies són inútils llevat que científics brillants plantegen preguntes rellevants i interpreten els resultats en el context apropiat.

Un mantra de l'estadística indica que la correlació no implica causalitat. Nathan Eagle es va avançar en la predicció del còlera a Rwanda a partir de les da-

Leaders (Executive Sport Ltd)

El relat de l'aplicació que el mànager de beisbol Billy Beane va fer dels indicadors de rendiment i d'anàlisi de dades per a formar un equip competitiu s'ha convertit ja en tot un clàssic contemporani de les històries d'èxit de l'aprofitament de les dades i fins i tot va donar lloc a una pel·lícula de Hollywood de bastant èxit.

des de mobilitat que va extraure de les cridades de telèfons mòbils (Shaw, 2014). Eagle va observar que la mobilitat estava correlacionada amb els brots de còlera i que, per tant, podia ajudar a predir-los. Després va descobrir que la mobilitat realment predeia les inundacions, que redueixen la mobilitat i incrementen a curt termini el risc de brots de còlera. Actualment incorpora informació sobre l'activitat de les poblacions en les seues prediccions. No hi ha res que puga reemplaçar la comprensió del fenomen que s'estudia, és a dir, el procés de generació de dades, per a poder analitzar-lo.

«LA TEORIA ENS ENSENYA QUE, EN PRINCIPI, TENIR MOLTES DADES SEMPRE ÉS BO. UNA TRAMPA TEMPTADORA CONSISTEIX A FORÇAR LES DADES FINS QUE SEMBLEN ABONAR UNA IDEA PRECONCEBUDA»

DINÀMICA DE DADES

Els Centres de Control i Prevenció de Malalties (CDC) dels EUA remeten setmanalment el nombre de visites mèdiques per malalties de tipus gripal, però els resultats van amb tres setmanes de retard, que és el que costa processar-los. Google Flu Trends (GFT) utilitza el nombre de cerques en Internet relacionades amb la grip per predir l'eventual informe dels CDC per a la setmana en curs, proporcionant un seguiment en temps real que es va considerar en alguna ocasió més precís que els informes dels mateixos CDC. Encara que GFT no ho pretenia, s'ha convertit en l'estendard del reemplaçament dels mètodes tradicionals per dades massives. No obstant això, Lazer *et al*. (2014), entre altres, han esbrinat que les prediccions de GFT no són tan precises. Encara que al començament sí que ho eren, des de llavors les visites reals sempre s'han sobreestimat. Predir simplement una setmana a partir dels informes dels CDC de tres setmanes arrere dóna millors resultats. Lazer *et al*. argumenten que la caiguda en la precisió de GFT és deguda sobretot als canvis en el motor de cerca de Google. Aquest exemple il·lustra una altre parany important. En el cas del beisbol i en la resta d'exemples anteriors, el procés subjacent que genera les dades sol romandre constant al llarg del temps. El beisbol té unes regles fixes, la intenció de vot no varia molt a la curta, i les lleis de la naturalesa són constants. Per contra, els canvis en els cercadors alteren el procés de generació de les dades que s'introdueixen en GFT i per consegüent modifiquen la relació amb el resultat que intentem predir.

Aquesta incertesa, en la literatura estadística, es coneix com a *sistema dinàmic* i requereix tècniques especials per a incorporar la seua peculiar estructura i poder reflectir la incertesa de manera fidedigna. Les prediccions es basen en les dades observades i, per tant, un supòsit implícit és que les dades futures seran semblants o almenys evolucionaran d'una manera previsible. Quan poden donar-se canvis sobtats, la confiança en les nostres prediccions disminueix. Considerem el fracàs a l'hora de preveure els impagaments d'hipoteques en la Gran Recessió. El risc d'impagament es calculava a partir de les dades aplegades durant un període de creixement econòmic generalitzat. En aquests períodes el risc que els individus A i B deixen de pagar les seues hipoteques no presenta cap correlació en particular. Per tant, el risc d'impagaments generalitzats es considera baix i encara que alguns individus deixen de pagar, segurament altres continuaran essent solvents. No obstant això, en períodes de crisi els impagaments estan estretament correlacionats. Si l'economia va malament i el preu de l'habitatge cau, molta gent es torna insolvent al mateix temps i les possibilitats d'una crisi generalitzada són molt majors (Gorton, 2009). Aquest exemple il·lustra un biaix conegut com a extrapolació. Fins i tot quan sabem una mica sobre el procés de generació de dades, és arriscat fer prediccions en situacions en què hi ha poques o cap dada disponibles. La majoria de mètodes estan dissenyats per a produir prediccions que en general són vàlides, però, encara que la majoria de les prediccions siguen necessàries, les que es desenvolupen en escenaris poc freqüents (per exemple, pacients amb una variant rara d'una malaltia) poden fallar completament. Cal, per tant, examinar acuradament el problema que ens ocupa.

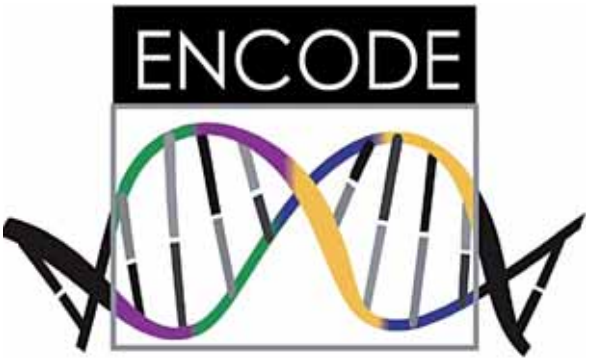
SENYAL, SOROLL I BIAIX

La teoria ens ensenya que, en principi, tenir moltes dades sempre és bo. Amb noves dades s'incrementa el potencial per a obtenir més informació i, si aquest no fóra el cas, sempre es podria descartar la dada. No sembla que tenir més dades faça nosa. L'error d'aquest raonament és que en la pràctica no descartem dades sinó que tractem de buscar-los algun patró. Una trampa temptadora consisteix a forçar les dades fins que semblen abonar una idea preconcebuda. Això no vol dir que l'anàlisi de dades no puga ser motivat per una hipòtesi prèvia, sinó que es necessita una estratègia adequada per a reduir la probabilitat d'obtenir-ne resultats no reproduïbles. Les últimes dècades han mostrat avenços apassionants en mètodes estadístics orientats a distingir el senyal del soroll en les dades massives. Però aquests avenços encara no han calat en les anàlisis rutinàries de dades. Nuzzo (2014) considera que, en observar un valor *p* de 0,01 per a una hipòtesi amb dinou probabilitats contra una de no ser certa, la probabilitat que es tracte d'un fals positiu és del 0,89. Amb les dades massives sovint es registren dades simplement perquè les podem obtenir, no perquè hom espere incrementar substancialment el senyal. Les probabilitats, per tant, són molt superiors a dinou contra una i les possibilitats de falsos positius es disparen.

Una altra qüestió determinant és que les dades massives sovint procedeixen de diferents fonts, s'han obtingut mitjançant diferents tècniques o presenten diferents formats. No necessàriament han de ser comparables o presentar la mateixa qualitat i sovint estan sotmeses a diversos errors sistemàtics. Per exemple, el projecte Encode és una de les majors iniciatives posteriors al Projecte Genoma Humà. Les dades es recopilen en laboratoris repartits pertot arreu, usant múltiples tecnologies i procediments experimentals. Quan desenvolupàrem un sistema per a visualitzar aquestes dades massives, trobàrem biaixos sistemàtics entre els microbioxips i les tecnologies de seqüenciació que s'havien de corregir per a evitar interpretacions errònies (Font-Burgada *et al*., 2013). Més en general, visualitzar dades heterogènies de manera que siguen fàcils d'interpretar és un repte, però s'hi van fent progressos. Per exemple, amb les tècniques de visualització del flux sanguini ideades per Michelle Borkin i els seus directors de tesi s'incrementava d'un 39 a un 91 % la capacitat dels metges per a diagnosticar obstruccions arterials (Shaw, 2014). En el passat, els mètodes de metaanàlisi es van concebre per combinar indicis de diferents estudis seguint un procediment rigorós. Les dades massives requereixen nous mètodes per a poder integrar i visualitzar les dades de manera fiable.

MÈTODE

Els Centres de Control i Prevenció de Malalties (CDC) dels EUA remeten setmanalment el nombre de visites mèdiques per malalties de tipus gripal, però els resultats van amb tres setmanes de retard, que és el que costa processar-los. Google Flu Trends (GFT) utilitza el nombre de cerques en Internet relacionades amb la grip per predir l'eventual informe dels CDC per a la setmana en curs, proporcionant un seguiment en temps real que ha estat considerat més precís que els informes dels mateixos CDC.



ENCODE PROJECT

Les dades massives sovint procedeixen de diferents llocs, s'han obtingut mitjançant diferents tècniques o presenten diferents formats. No sempre són comparables o ofereixen la mateixa qualitat i a vegades estan sotmeses a diversos biaixos sistemàtics. Aquest és el tipus de problemes que experimenta Encode, una de les iniciatives més importants posteriors al Projecte Genoma Humà. Les dades es van recopilar en laboratoris repartits per tot el món, usant múltiples tecnologies i procediments experimentals.

PLANIFICAR

Les dades massives estan canviant la manera de recopilar resultats. En compte de dissenyar acuradament un estudi, la tendència sol ser registrar totes les dades que siga possible, acceptant de manera implícita que qualsevol patró que s'hi observe segurament serà rellevant. Aquesta idea falsa és una trampa molt problemàtica. La representativitat de les dades no depèn de la grandària de la mostra sinó de la manera de recopilar-les. Importa més la qualitat que la quantitat. Un exemple clàssic és un estudi britànic en què es van avaluar en 20.000 nens els beneficis de la llet pasteuritzada. William Gosset, més conegut com Student, va assenyalar que, per culpa d'una distribució aleatòria inadequada, un estudi amb només sis bessons hauria estat més fiable (Student, 1931). Un factor que contribueix a descuidar el disseny de l'estudi pot ser l'excés de fe en les noves tecnologies. Per exemple, la comunitat científica ha rebut amb entusiasme la irrupció de la seqüenciació d'alt rendiment (HTS). He conegut reputats investigadors que argumenten que amb una sola mostra aquests estudis són tan bons com les tecnologies anteriors amb dotzenes de mostres. Encara que l'HTS siga necessària, una sola mostra no pot mesurar la variabilitat quan es compararen universos. Una altra anècdota és que alguns centres d'HTS processen dues mostres en diferents dates quan haurien d'haver-les processades en paral·lel per evitar biaixos. Com a resultat, experiments molt cars han donat resultats pràcticament inútils.

L'extensió de la teoria sobre el disseny d'experiments formulada per Ronald Fisher a les dades massives ha caigut quasi en l'oblit, però hi ha notables excepcions. D'acord amb la tendència cap a la medicina personalitzada, Berry (2012) ha defensat els assajos clínics adaptats a grups cada vegada més reduïts i a la presa de decisions per a cada pacient. Müller *et al*. (2004) han proposat dissenys rigorosos per a estudis de comprovació massiva d'hipòtesis. També s'han proposat dissenys reeixits d'estudis observacionals. Per a mostrar els avantatges de l'assegurança mèdica pública a Mèxic, King *et al*. (2009) van elaborar un estudi que comparava les comunitats que tenen aquesta assegurança i les que no en tenien. Com que aquestes mostraven característiques semblants, les diferències entre els resultats en salut es poden atribuir més aviat a l'assegurança que no a factors externs.

MÈTODE

Els suggeriments de pel·lícules que fa Netflix utilitzen un model que fa la mitjana de 107 prediccions. La teoria de la decisió pot ajudar a avaluar els avantatges d'algoritmes complexos en un context dominat per la incertesa i els objectius contraposats; per exemple, el grau de satisfacció dels clients també pot dependre de la diversitat dels suggeriments.

UN CAS PER A L'ESTADÍSTICA

Pioners com Ronald Fisher, William Gosset o Harold Jeffreys van establir les bases de l'aplicació de les dades a la ciència, els negocis i la política. D'una manera semblant, el paradigma de les dades massives s'ha alimentat de contribucions metodològiques. L'algoritme PageRank utilitzat per Google es basa en les cadenes de Markov. Els suggeriments de pel·lícules que fa Netflix utilitzen un model que fa la mitjana de 107 prediccions. La teoria de la decisió pot ajudar a avaluar els avantatges de complexos algoritmes en un context dominat per la incertesa i per objectius contraposats; per exemple, el grau de satisfacció dels clients també pot dependre de la diversitat dels suggeriments.

Ja hem exposat la necessitat d'explorar nous mètodes per a destriar el senyal del soroll, capturar processos dinàmics, dissenyar experiments i integrar dades heterogènies. Els mètodes computacionals que combinen potència de processament amb estratègies intel·ligents per a resoldre problemes complexos són un altre dels temes decisius, ja que és poc probable que reïsquen els enfocaments exhaustius o de força bruta. Altres reptes són la recuperació i el resum de dades. Els mètodes automàtics per a escanejar i donar format a les dades no estructurades (com ara imatges o

«EN COMPTE DE DISSENYAR ACURADAMENT UN ESTUDI, LA TENDÈNCIA SOL SER REGISTRAR TOTES LES DADES QUE SIGA POSSIBLE, ACCEPTANT QUE QUALSEVOL PATRÓ QUE S'HI OBSERVE SEGURAMENT SERÀ RELLEVANT»

blogs) poden bandejar informació o induir biaixos. Un altre problema és que actualment generem més dades que no les que podem emmagatzemar (Hilbert, 2012), el que obliga a resumir-les. I els resums impliquen el risc de perdre informació. Com a exemple, fa poc vam explicar que l'estratègia que actualment s'aplica per a recapitular les dades de la seqüenciació de ARN descarta tanta informació que certs detalls s'escapen encara que la quantitat de dades vaja creixent fins a l'infinit (Rossell *et al*., 2014). Un tema relacionat és el de la presa de mostres. Emmagatzemar una mostra apropiada obtinguda de totes les dades pot incrementar la velocitat i reduir costos, amb una pèrdua insignificant en la precisió. Fan *et al*. (2014) i Jordan (2013) han abordat qüestions relatives a l'estadística i al processament de dades massives.

L'estadística, com a disciplina que combina raonament científic, teoria de la probabilitat i matemàtiques, és un component necessari perquè la revolució de les dades massives assolesca tot el seu potencial. No obstant això l'estadística no pot funcionar de manera aïllada sinó que necessita la col·laboració de coneixements tècnics, de la informàtica i d'altres disciplines relacionades. Com a reflexió final, el principal obstacle que cal superar ben bé pot ser la falta de professionals amb la combinació adequada de capacitats. La selecció i la formació de joves talents disposats a participar en aquesta excitant aventura hauria de ser una prioritat.

«L'ESTADÍSTICA ÉS UN COMPONENT NECESSARI PERQUÈ LA REVOLUCIÓ DE LES DADES MASSIVES ASSOLESCA TOT EL SEU POTENCIAL PERÒ NO POT FUNCIONAR DE MANERA AÏLLADA»

REFERÈNCIES

Berry, D., 2012. «Adaptive Clinical Trials in Oncology». *Nature Reviews Clinical Oncology,* 9: 199-207. DOI: <10.1038/nrclinonc.2011.165>.

Curtice, J. i D. Firth, 2008. «Exit Polling in a Cold Climate: the BBC-ITV Experience Explained». *Journal of the Royal Statistical Society A,* 171(3): 509-539. DOI: <10.1111/j.1467-985X.2007.00536.x>.

Fan, J.; Han, F. i H. Liu, 2014. «Challenges of Big Data Analysis». *National Science Review*, 1(2): 293-314. DOI: <10.1093/nsr/nwt032>.

Font-Burgada, J.; Reina, O.; Rossell, D. i F. Azorín, 2013. «ChroGPS, a Global Chromatin Positioning System for the Functional Analysis and Visualization of the Epigenome». *Nucleic Acids Research,* 42(4): 1-12. DOI: <10.1093/nar/gkt1186>.

Fòrum Econòmic Mundial, 2012. *Big Data, Big Impact: New Possibilities for International Development*. Fòrum Econòmic Mundial. Cologny, Suïssa. Disponible en: <www3.weforum.org/docs/WEF_TC_MFS_BigData-BigImpact_Briefing_2012.pdf>.

Gorton, G., 2009. «Information, Liquidity, and the (Ongoing) Panic of 2007». *American Economic Review*, 99(2): 567-572. DOI: <10.1257/aer.99.2.567>.

Hilbert, M., 2012. «How Much Information Is There in the "Information Society"?». *Significance,* 9(4): 8-12. DOI: <10.1111/j.1740-9713.2012.00584.x>.

International Business Machines Corporation, 2011. *IBM Big Data Success Stories*. International Business Machines Corporation. Armonk, NY. Disponible en: <http://public.dhe.ibm.com/software/data/sw-library/big-data/ibm-big-data-success.pdf>.

Jordan, M., 2013. «On Statistics, Computation and Scalability». *Bernoulli*, 19(4): 1378-1390. DOI: <10.3150/12-BEJSP17>.

King, G. *et al*., 2009. «Public Policy for the Poor? A Randomized Assessment of the Mexican Universal Health Insurance Programme». *The Lancet,* 373: 1447-1454. DOI: <10.1016/S0140-6736(09)60239-7>.

Lazer, D.; Kennedy, R.; King, G. i A. Vespignani, 2014. «The Parable of Google Flu: Traps in Big Data Analysis». *Science*, 343(6176): 1203-1205. DOI: <10.1126/science.1248506>.

Lewis, M., 2003. *Moneyball. The Art of Winning an Unfair Game*. W. W. Norton & Company. Nova York.

Lohr, S., 2012. «The age of Big Data». *The New York Times*, 11 de febrer de 2012. Disponible en: <www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html>.

Müller, P.; Parmigiani, G.; Robert, C. i J. Rousseau, 2004. «Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays». *Journal of the American Statistical Association,* 99(468): 990-1001. DOI: <10.1198/016214504000001646>.

Nuzzo, R., 2014. «Scientific Method: Statistical Errors». *Nature*, 506: 150-152. DOI:<10.1038/506150a>.

Rossell, D.; Stephan-Otto Attolini, C.; Kroiss, M. i A. Stöcker, 2014. «Quantifying Alternative Splicing from RNA-Sequencing Data». *The Annals of Applied Statistics*, 8(1): 309-330. DOI: <10.1214/13-AOAS687>.

Shaw, J., 2014. «Why "Big Data" Is a Big Deal». *Harvard Magazine*, 3: 30-35, 74-75. Disponible en: <http://harvardmag.com/pdf/2014/03-pdfs/0314-30.pdf>.

Silver, N., 2012. *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't*. Penguin Press. Nova York.

Student, 1931. «The Lanarkshire Milk Experiment». *Biometrika*, 23(3-4): 398-406. DOI: <10.2307/2332424>.

ABSTRACT

Big Data and Statistics: A Statistician's Perspective.

Big Data brings unprecedented power to address scientific, economic and societal issues, but also amplifies the possibility of certain pitfalls. These include using purely data-driven approaches that disregard understanding the phenomenon under study, aiming at a dynamically moving target, ignoring critical data-collection issues, summarizing or preprocessing the data inadequately and mistaking noise for signal. We review some success stories and illustrate how statistical principles can help obtain more reliable information from data. We also touch upon current challenges that require active methodological research such as strategies for efficient computation, integration of heterogeneous data, extending the underlying theory to increasingly complex questions and, perhaps most importantly, training a new generation of scientists who can develop and deploy these strategies.

Keywords: Big Data, statistics, case studies, pitfalls, challenges.

AGRAÏMENTS:
Treball parcialment finançat per NIH grant R01 CA158113-01.

David Rossell. Professor del departament d'Estadística. Universitat de Warwick (Regne Unit).