

Prezenta ediție românească este publicată prin acordul cu Houghton Mifflin Harcourt Publishing Company prin Agenția Literară Livia Stoia.

Big Data: A Revolution That Will Transform How We Live, Work, and Think

Copyright © 2013 by Viktor Mayer-Schönberger and Kenneth Cukier

Originally published by Houghton Mifflin Harcourt, 2013.

© 2018 Editura ACT și Politon pentru prezenta ediție românească

Editura ACT și Politon

Str. Înclinată, nr. 129, Sector 5, București, România, C.P. 050202.

tel: 0723 150 590, e-mail: office@actsipoliton.ro

www.actsipoliton.ro/blog

Traducător: **Romică Lixandru**

Redactor: **Ines Simionescu**

Editor: **Camelia Zara**

Tehnoredactor: **Teodora Vlădescu**

Coperta: **Marian Iordache**

Copyright Manager: **Andrei Popa**

Descrierea CIP a Bibliotecii Naționale a României

MAYER-SCHONBERGER, VIKTOR

Big Data / Viktor Mayer-Schönberger & Kenneth Cukier ; trad.:

Romică Lixandru. - București : ACT și Politon, 2018

ISBN 978-606-913-401-6

I. Cukier, Kenneth

II. Lixandru, Romică (trad.)

004

AVERTISMENT: Distribuirea, copierea sau piratarea în orice fel a acestei cărți nu este pedepsită numai prin lege, dar contravine și tuturor normelor și principiilor etice și sănătoase pe care un astfel de titlu le promovează. Ce fel de efect va avea energia pe care vrei să o transmiți mai departe, dacă aceasta vine prin furt, ilegalitate și lipsă de respect față de autor și față de toți cei care au contribuit la crearea acestei cărți, astfel ca ea să ajungă la dumneavoastră? Împărtășiți cu ceilalți informațiile importante, valorile și lecțiile pe care le-ați aflat din acest material, într-un mod corect și responsabil.

**VIKTOR MAYER-SCHÖNBERGER
ȘI KENNETH CUKIER**

BIG DATA

**O revoluție care va transforma felul
în care trăim, muncim și gândim**

Traducere din limba engleză de
Romică Lixandru

 **ACT și Politon**

2018

Pentru B și v

V.M.S.

Părinților mei

K.N.C.

Cuprins

Capitolul 1:	Acum	7
Capitolul 2:	Mai mult	33
Capitolul 3:	În dezordine	52
Capitolul 4:	Corelații	76
Capitolul 5:	„Dataficarea”	108
Capitolul 6:	Valoarea	143
Capitolul 7:	Implicațiile	179
Capitolul 8:	Riscuri	218
Capitolul 9:	Controlul	247
Capitolul 10:	Ce urmează	267
	<i>Mulțumiri</i>	286
	<i>Note</i>	290
	<i>Bibliografie</i>	309

1

ACUM

ÎN 2009 A FOST DESCOPERIT un nou virus gripal. Cu o combinație de elemente ale virusurilor care provoacă gripa aviară și gripa porcină, acest nou virus H1N1 s-a răspândit rapid. În decurs de câteva săptămâni, institutele de sănătate publică din toată lumea se temeau că se va declanșa o pandemie îngrozitoare. Câțiva comentatori avertizau că este posibilă o epidemie de dimensiunile gripei spaniole din 1918 care a afectat o jumătate de miliard de oameni și a omorât zeci de milioane. Cel mai rău era că nu exista niciun vaccin să poată fi utilizat imediat împotriva noului virus. Singura speranță pe care o aveau autoritățile din domeniul sănătății publice era aceea de a-i încetini răspândirea. Dar, pentru asta, trebuiau să determine unde se găsea virusul.

În Statele Unite, conducerea Centrului pentru Controlul și Prevenirea Bolilor (Centers for Disease Control and Prevention) a cerut medicilor să raporteze cazurile noi de gripă. Cu toate acestea, tabloul obținut era întotdeauna cu una sau două săptămâni în urmă. În plus, deși oamenii se simțeau rău zile la rând, nu se duceau imediat să consulte un medic. În afară de asta, transmiterea acelor informații

Respectiv, către Centru lua ceva timp. Organizația înregistra cifrele doar o dată pe săptămână. În cazul unei boli care se răspândește rapid, un decalaj de două săptămâni este o eternitate. Această întârziere a lăsat în necunoștință totală de cauză agențiile de sănătate publică exact în cele mai importante momente.

Din întâmplare, cu câteva săptămâni înainte ca virusul H1N1 să ajungă pe prima pagină a ziarelor, inginerii de la compania gigant din domeniul Internetului, Google, au publicat un articol remarcabil în revista științifică *Nature*¹. În rândul autorităților din domeniul sănătății și al informaticienilor a avut un impact semnificativ, dar în rest a fost trecut cu vederea. Acesta explica felul în care Google putea să „prezică” răspândirea gripei din timpul iernii în Statele Unite nu doar la nivel național, dar și regional și chiar statal. Pentru a reuși acest lucru, Google a analizat ce anume căutau oamenii pe Internet. Cum existau peste trei miliarde de căutări zilnice la nivel global, Google avea o mulțime de date cu care să lucreze. A ajutat însă și faptul că stochează toate cererile de căutare pe care le primește timp de mai mulți ani.

Google a luat primii 50 de milioane de termeni cel mai des întâlniți din căutările făcute de americani și i-a comparat cu datele Centrului despre răspândirea gripei sezoniere între 2003 și 2008. Ideea era că oamenii infectați cu virusul gripei ar putea fi identificați după căutările pe care le fac pe Internet. Încercaseră și alții să stabilească acest lucru cu exactitate din datele despre căutările pe Internet, dar nimeni nu avea la fel de multe informații, la fel de multă putere de procesare și la fel de multă cunoaștere în domeniul statisticii precum Google.

Chiar dacă inginerii de la Google bănuiau că acele căutări erau făcute cu scopul de a obține informații referitoare

la gripă – fraze cum ar fi „medicamente pentru tuse și febră” –, ideea era alta: nu știau sigur lucrul acesta și au proiectat un sistem care să nu țină cont de asta. Sistemul lor nu făcea altceva decât să caute corelații între frecvența anumitor căutări și răspândirea gripei în timp și spațiu. Google a procesat în total un număr uluitor de 450 de milioane de modele matematice diferite pentru a testa termenii de căutare, comparând predicțiile acestora cu cazurile reale de gripă de la Centrul pentru Controlul și Prevenirea Bolilor din 2007 și 2008. Inginerii de la Google au tras lozul cel mare: softul lor a găsit o combinație de 45 de termeni care, atunci când erau utilizați împreună într-un model matematic, creau o corelație de 97% între predicția lor și datele oficiale. Asemenea Centrului, ei știau unde se răspândise gripa, dar, spre deosebire de acesta, aflau aproape în timp real, nu după o săptămână sau două.

Prin urmare, când s-a declanșat criza H1N1 din 2009, sistemul s-a dovedit a fi un indicator mult mai util și mai oportun decât statisticile oficiale ale guvernului, cu decalajele lor de raportare². Autoritățile din domeniul sănătății publice erau înarmate cu informații valoroase. În mod surprinzător, metoda Google nu presupune distribuirea de bețișoare de prelevare sau contactarea cabinetelor medicilor. Dimpotrivă, era construită pe „big data” – capacitatea societății de a utiliza informațiile în feluri noi pentru a obține perspective utile sau bune și servicii de valoare semnificativă. Cu ajutorul lor, atunci când își va face apariția următoarea pandemie, lumea va avea la dispoziție un instrument mai bun prin care să o prevadă și, astfel, să prevină răspândirea ei.

* Seturi de date care sunt atât de mari și complexe, încât aplicațiile de software tradiționale nu reușesc să le proceseze. (n.tr.)

Sănătatea publică este doar unul dintre domeniile în care big data fac o mare diferență. Sectoare întregi ale activității economice sunt de asemenea remodelate de big data. Un bun exemplu sunt biletele de avion.³

În 2003, Oren Etzioni trebuia să ajungă cu avionul din Seattle la Los Angeles pentru nunta fratelui său mai mic. Cu câteva luni înainte de ziua cea mare, a intrat online și a cumpărat un bilet de avion, știind că cu cât îl rezervi mai din timp, cu atât plătești mai puțin. În timpul zborului, totuși, nu și-a putut stăpâni curiozitatea și l-a întrebat pe cel de lângă el cât plătise biletul. Dăduse mult mai puțin, deși îl cumpărase mult mai recent. Înfuriat, Etzioni a mai întrebat un pasager și apoi un altul. Toți dăduseră mai puțin.

Pentru majoritatea oamenilor, sentimentul de a fi fost tras pe sfoară s-ar fi risipit până la momentul în care trebuie să ridice măsuțele și să-și așeze scaunele în poziție verticală. Dar Etzioni este unul dintre informaticienii de vârf ai Americii. Din postura sa foarte avantajoasă de director al programului de inteligență artificială de la Universitatea din Washington, el a înființat o mulțime de companii de big data înainte ca sintagma „big data” să devină cunoscută.

A contribuit la construirea unuia dintre primele motoare de căutare de pe internet în 1995, numit MetaCrawler, care a fost preluat de Infospace când a devenit o proprietate online importantă. A fost unul dintre cofondatorii Netbot, primul program semnificativ de comparare a prețurilor, pe care l-a vândut firmei Excite. Clearforest, cum se numea start-up-ul său specializat în extragerea de informații din documentele text, avea să fie preluat mai târziu de Reuters. Etzioni privește lumea ca pe o uriașă problemă de informatică – una pe care o poate rezolva. Și a rezolvat multe

dintre ele de când a absolvit la Harvard în 1986, unde a fost primul student care a obținut o specializare în informatică.

Ajuns la destinație, Etzioni era hotărât să găsească o modalitate prin care oamenii să afle dacă prețul unui bilet de avion pe care îl găsesc online reprezintă o afacere bună sau nu. Locurile dintr-un avion sunt bunuri: fiecare dintre ele este, în principiu, imposibil de deosebit de celelalte din aceeași cursă. Și totuși, prețurile diferă foarte mult în funcție de o mulțime de factori cunoscuți doar de companiile de transport aerian.

Etzioni a ajuns la concluzia că nu trebuia să descifreze logica sau motivele diferențelor de preț, ci doar să anticipeze dacă prețul afișat era mai probabil să crească sau să scadă. Este posibil să afli acest lucru, ba chiar ușor. Tot ce trebuie să faci este să analizezi prețurile biletelor pentru o anumită rută în funcție de numărul zilelor rămase până la plecare.

Dacă prețul mediu al unui bilet are tendința să scadă, ar fi logic să aștepti și să cumperi biletul mai târziu. Dacă prețul mediu are tendința să crească, sistemul recomandă cumpărarea imediată a biletului la prețul afișat. Cu alte cuvinte, era o versiune elaborată a sondajului neoficial efectuat de Etzioni la 9.000 de metri altitudine. Bineînțeles, era o problemă uriașă de informatică, dar una pe care putea să o rezolve. Așa că s-a apucat de treabă.

Folosind un eșantion alcătuit din 12.000 de examinări ale prețurilor de-a lungul unei perioade de 41 de zile, prețuri obținute prin „extragerea” informațiilor de pe un site de călătorii, Etzioni a creat un model predictiv care le aducea pasagerilor săi fictivi economii bune. Sistemul nu știa să spună *de ce*, ci doar *cât*. Adică nu cunoștea niciuna dintre variabilele care influențează deciziile prin care

Respe companiile de transport aerian stabilesc prețul, cum ar fi numărul de locuri care nu au fost vândute, caracterul sezonier sau dacă exista vreun fel de „cazare magică în noaptea de sâmbătă” ce putea să reducă tariful. Știa doar să-și facă predicțiile bazându-se pe probabilitățile rezultate din utilizarea datelor despre alte zboruri. „A cumpăra sau a nu cumpăra, aceasta-i întrebarea”, a cugetat Etzioni. În mod foarte potrivit, el și-a numit proiectul de cercetare Hamlet.⁴

Micul proiect a evoluat într-un start-up numit Farecast*, finanțat cu capital de risc. Prin faptul că prognoza dacă prețul unui bilet de avion era mai probabil să crească sau să scadă, precum și cu cât anume, le dădea consumatorilor posibilitatea să aleagă când să achiziționeze. El i-a înarmat cu informații la care nu au avut niciodată acces. Aplicând calitatea transparenței la sine însuși, Farecast a estimat gradul de încredere al propriilor predicții și a prezentat și această informație.

Sistemul avea nevoie de o mulțime de date pentru a funcționa. Pentru a-i îmbunătăți performanța, Etzioni a reușit să intre în posesia unei baze de date cu rezervări de bilete de avion. Cu aceste informații, sistemul putea să facă predicții bazate pe toate locurile din toate zborurile de pe toate rutele din aviația comercială americană de-a lungul unui an. Pentru calcularea predicțiilor, Farecast prelucra acum aproape 200 de miliarde de date legate de zboruri. În felul acesta îi ajuta pe consumatori să economisească o mulțime de bani.

Cu părul șaten și ciufulit, zâmbetul larg și frumusețea angelică, Etzioni nu părea genul de persoană care să refuze industriei aviatice accesul la venituri potențiale de milioane de dolari. De fapt, el avea scopuri și mai mari. Etzioni

* De la *fare* = tarif, și *cast* = prognoză. (n.ed.)

plănuia ca până în 2008 să aplice metoda și la alte lucruri, cum ar fi camerele de hotel, biletele la concerte și mașinile la mâna a doua: orice domeniu unde produsele se diferențiază prea puțin, unde există un grad ridicat de variație a prețului și tone de date. Dar înainte să-și poată elabora planurile, Microsoft i-a bătut la ușă și a cumpărat rapid Farecast cu 110 milioane de dolari și l-a integrat în motorul de căutare Bing.⁵ Până în 2012, sistemul ajunsese să analizeze în jur de un trilion de intrări pe baza cărora a făcut predicții pentru tarifele tuturor zborurilor interne din Statele Unite. A făcut predicții corecte în 75% din cazuri și i-a ajutat pe călători să economisească, în medie, 50 \$ la fiecare bilet.⁶

Farecast este un model al companiilor care lucrează cu big data și un exemplu care ilustrează direcția în care se îndreaptă lumea. Etzioni nu ar fi putut să înființeze compania cu cinci sau cu zece ani mai devreme. „Ar fi fost imposibil”, spune el. Puterea de calcul și spațiul de stocare de care avea nevoie erau prea costisitoare. Dar chiar dacă schimbările tehnologice au fost un factor deosebit de important care a făcut posibilă apariția companiei, a mai existat ceva care s-a schimbat, ceva subtil și chiar și mai important: mentalitatea cu privire la felul în care puteau fi utilizate datele.

Datele nu mai erau privite ca obiecte statice sau depășite, a căror utilitate se încheie odată cu atingerea obiectivului pentru care fuseseră colectate, cum ar fi după aterizarea avionului (sau, în cazul Google, după procesarea unei căutări). Datele au devenit mai degrabă un material brut pentru afaceri, o informație economică vitală utilizată pentru a crea o nouă formă de valoare economică. De fapt, dacă există o mentalitate adecvată, datele pot fi reutilizate

Respe în mod inteligent pentru a deveni un izvor de inovație și de noi servicii. Datele pot dezvălui secrete celor care au modestia, deschiderea și mijloacele să asculte.

SĂ LĂSĂM DATELE SĂ VORBEASCĂ

Roadele societății informaționale sunt ușor de văzut atunci când există câte un telefon mobil în fiecare buzunar, câte un computer pe fiecare birou și sisteme mari de IT în departamentele administrative de peste tot. Pe de altă parte, informația în sine este mai puțin vizibilă. La o jumătate de secol după ce computerele au intrat în forță în societate, datele au început să se acumuleze până în punctul în care începe să se petreacă un lucru nou și deosebit. Nu numai că lumea este inundată de informații mai mult ca oricând, dar acele informații devin din ce în ce mai numeroase. Schimbarea proporțiilor a dus la o schimbare a stării. Schimbarea cantitativă a generat o schimbare calitativă. Științe precum genomica și astronomia, primele care au resimțit explozia de la jumătatea anilor 2000, au inventat sintagma „big data”. Conceptul migrează acum în toate domeniile activității umane.

Nu există nicio definiție riguroasă a big data. Inițial, ideea a fost că volumul de informații a devenit atât de mare, încât nu mai încăpea efectiv în memoria pe care computerul o utilizează pentru procesare, astfel că inginerii trebuiau să modernizeze instrumentele pe care le utilizau pentru analizarea lor. Aceasta este originea noilor tehnologii de procesare cum ar fi MapReduce de la Google și a programului echivalent open source, Hadoop, lansat de Yahoo. Ele ne dau posibilitatea să gestionăm cantități de date mult mai mari decât înainte și care – un lucru important – nu trebuie să fie aranjate în rânduri sau tabelele clasice. La

orizont se prefigurează și alte tehnologii pentru analizarea datelor care se dispensează de ierarhiile rigide și de omogenitatea de odinioară. În același timp, din moment ce companiile din industria Internetului puteau să colecteze cantități industriale de date și aveau un interes financiar imperios să le înțeleagă, au devenit utilizatorii de top ai tehnologiilor de procesare de ultimă oră, depășind companii care aveau câteva decenii în plus de experiență în unele cazuri, dar își desfășurau activitatea offline.

În ziua de azi, ne putem gândi la această chestiune în felul următor (așa cum l-am adoptat și în această carte): big data se referă la lucruri pe care le putem face la scară mare, dar care nu pot fi făcute la o scară mai mică, pentru a obține noi perspective sau a crea noi forme de valoare, într-un mod care schimbă piețele, organizațiile, relația dintre cetățeni și guverne și multe altele.

Dar acesta este doar începutul. Era big data pune sub semnul întrebării felul în care trăim și interacționăm cu lumea. Lucrul cel mai surprinzător este că societatea va trebui să renunțe la căutarea cauzalității în schimbul unor corelații simple: nu va ști *de ce*, ci doar *ce*. Lucrul acesta răstoarnă secole de practică împământenită și pune la încercare înțelegerea fundamentală a felului în care luăm decizii și în care percepem realitatea.

Big data marchează începutul unei transformări majore. Așa cum telescopul ne-a oferit posibilitatea să înțelegem universul iar microscopul ne-a permis să înțelegem microbii, tehnicile noi de colectare și analizare a unui volum uriaș de date ne vor ajuta să înțelegem lumea în feluri pe care abia începem să le întrezărim. Și, din nou, adevărata revoluție nu constă în dispozitivele care calculează datele, ci în datele însăși și în felul în care le utilizăm.

Pentru a înțelege în ce măsură avem de-a face cu o revoluție informațională deja în desfășurare, gândește-te la tendințele din spectrul societății. Universul nostru digital se extinde constant. Să luăm cazul astronomiei. Când Sloan Digital Sky Survey* și-a început activitatea în anul 2000, telescopul lor din New Mexico a colectat în primele săptămâni mai multe date decât se acumulaseră în toată istoria astronomiei. În 2010, arhiva lor a ajuns la un volum uimitor de 140 de terabiți de informații. Dar un succesor, The Large Synoptic Survey Telescope, care urmează să intre în funcțiune în 2016 în Chile, va colecta această cantitate de date la fiecare cinci zile.⁷

Astfel de cantități astronomice se găsesc și ceva mai aproape de casă. Atunci când a fost descifrat genomul uman în 2003, a fost nevoie de un deceniu de muncă intensă pentru a obține secvența celor trei miliarde de perechi de bază. Aproape un deceniu mai târziu, exact aceeași cantitate de ADN este analizată la aproximativ fiecare 15 minute de aparatele genomice din întreaga lume.⁸ Costul pentru secvențierea genomului a scăzut sub 1.000 \$ în 2012, transformând-o într-o tehnică de masă care poate fi realizată la scară.⁹ În finanțe, pe piețele americane de capital sunt tranzacționate zilnic în jur de șapte miliarde de acțiuni, iar aproximativ două treimi din tranzacții sunt decise de algoritmi computerizați pe bază de modele matematice care calculează munți de date pentru a prezice profiturile în timp ce încearcă să reducă riscul.

În special companiile din domeniul internetului au fost luate cu asalt. Google procesează zilnic peste un petabit de date, un volum care înseamnă de aproximativ 100 de ori cantitatea de materiale tipărite din Biblioteca Congresului

*Program de cercetare și supraveghere a obiectelor cerești. (n.tr.)

Statelor Unite.¹⁰ Pe Facebook, o companie care nu exista în urmă cu un deceniu, sunt încărcate peste 10 milioane de fotografii pe oră. Oamenii dau clic pe „Îmi place” sau postează comentarii de aproape trei miliarde de ori pe zi, lăsând o urmă digitală pe care compania o poate analiza pentru a afla preferințele utilizatorilor.¹¹ Între timp, cei 800 de milioane de utilizatori lunari ai serviciului YouTube, oferit de Google, încarcă peste o oră de înregistrări video pe secundă.¹² Numărul mesajelor de pe Twitter crește cu aproximativ 200% pe an, iar până în 2012 a depășit 400 de milioane de tweet-uri pe zi.¹³

De la știință la sănătate, de la banking la Internet, chiar dacă domeniile sunt diverse, toate sugerează același lucru: cantitatea de date la nivel mondial crește rapid, depășind nu doar aparatura, ci și imaginația noastră.

Multe persoane au încercat să estimeze printr-o cifră concretă cantitatea de informații care ne înconjoară și să calculeze cât de repede crește. Gradul de succes obținut a variat pentru că măsurau lucruri diferite. Unul dintre cele mai cuprinzătoare studii a fost efectuat de Martin Hilbert de la Școala de Comunicații Annenberg a Universității din California de Sud.¹⁴ El a încercat să determine o cifră pentru tot ce a fost produs, stocat și comunicat. Aceasta ar include nu doar cărțile, picturile, e-mailurile, fotografiile, muzica și videoclipurile (analogice sau digitale), ci și jocurile video, apelurile telefonice, chiar și sistemele de navigare ale mașinilor și scrisorile trimise prin poștă. A inclus și televiziunea și radioul, în funcție de publicul la care ajunge.

După estimările sale, în 2007 au fost stocați sau difuzați aproximativ 2,25 zetabiți de date. Comparativ cu cei aproximativ 435 de exabiți de date de acum două decenii