

LBRIS

We know
books

TEORIA INFORMAȚIEI

Matematica în era digitală

JUAN M.R. PARRONDO

Traducere de Antonia Călin

LITERA
București

CUPRINS

Capitolul 0

Claude E. Shannon: jonglerii pe holurile din Bell Labs	7
--	---

Capitolul 1

Bazele teoriei informației	15
Sens, informație și aleatoritate	15
Întrebări și răspunsuri	17
<i>De la bit la qbit</i>	22
Informație și incertitudine	25
<i>Câte informații oferă o întrebare?</i>	26
Secvențe tipice	28
Entropia limbajului	31
<i>n-grame: modele aleatorii de limbaj</i>	34

Capitolul 2

Date, modele și complexitate	39
Coduri	39
<i>Codul lui Huffman</i>	44
Compresoare de fișiere	46
<i>Algoritmul Lempel-Ziv</i>	52
Fabrica de aleatoritate	52
Compresia finală: complexitatea algoritmică	59

Capitolul 3

Teoria comunicației	63
Canale, zgomot și remedierea erorilor	63

Entropie condiționată și informația reciprocă	71
<i>Matematica entropiei condiționate</i>	76
A doua teoremă a lui Shannon	78
Comunicare analogică	85
Lățime de bandă, semnale și zgomote	89
Comunicații fără fir	95
<i>Decibeli</i>	96

Capitolul 4

Informația în fizică	101
Entropia și axa timpului	101
Demonul lui Maxwell	108
Informația cuantică	113
Calculatoare cuantice	120
<i>Operatorul lui Grover</i>	128
Epilog: ce este informația?	130

Anexe	135
1. Proprietăți de bază ale logaritmilor	137
2. Sistemul binar	138

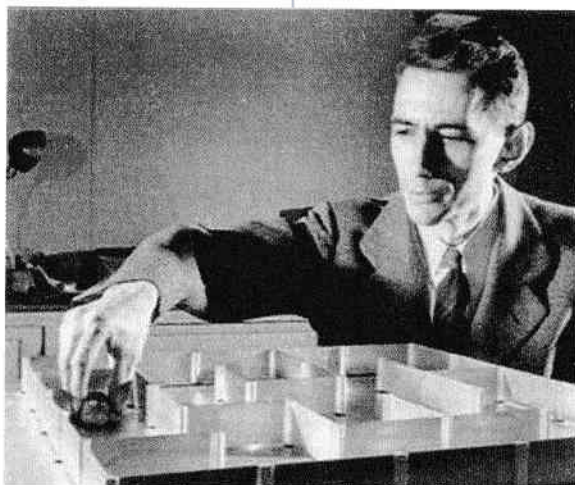
Bibliografie recomandată	141
---------------------------------	-----

Capitolul 0

Claude E. Shannon: jonglerii pe holurile din Bell Labs

Există anumite locuri mitice pentru știința și pentru tehnologia secolului XX: drumurile din Princeton, unde Einstein și Gödel probabil au făcut o plimbare împreună, Institutul de Fizică Teoretică din Copenhaga, unde s-a dezvoltat o mare parte din mecanica cuantică, hangarul din centrul Parisului, unde Marie și Pierre Curie au descoperit radioactivitatea... Unul dintre aceste locuri este, fără îndoială, Bell Labs, Laboratoarele Bell ale companiei ATT (acum, Nokia), unde s-a demonstrat natura undelor materiei (1927) și unde sau fost dezvoltate primul tranzistor (1947), prima celulă fotovoltaică (1954), primul senzor CCD (1969), limbajul Unix (1969) și s-au adus multe alte contribuții fundamentale la știință și la tehnologie. Este dificil să ne imaginăm o astfel de acumulare de talent și de creativitate într-o singură instituție.

La mijlocul secolului trecut și în mijlocul atâtor activități la cel mai înalt nivel în inginerie, în fizică și în matematică, un tip subțire și oarecum introvertit se plimba prin campusul Bell Labs, jonglând cu un monociclu: **Claude Elwood Shannon** (1916-2001), tatăl teoriei informației. Dacă astăzi suntem cufundați în societatea informațională, rezultatul așa-numitei a patra revoluție industrială se datorează în mare măsură contribuției acestui matematician, care constă, practic,



Claude Shannon în Bell Labs, în jurul anului 1952, cu Tezeu, șoarecele său, capabil să rezolve labirinturi.

în două articole publicate în 1948. În aceste două lucrări, Shannon a reușit să conceapă pentru prima dată o măsură cantitativă a informațiilor și să demonstreze două teoreme-cheie care caracterizează, cât de mult poate fi comprimat un set de date și cât de multe informații fiabile pot fi trimise printr-un canal de transmisie precum undele electromagnetice de la Wi-Fi sau prin cablurile telefonice de acasă.

Cei de la Bell Labs au realizat o excelentă expoziție virtuală, dedicată lui Claude Shannon, care poate fi vizitată pe linkul www.bell-labs.com/claude-shannon. Conține multe documente revelatoare despre activitatea lor și despre cea a laboratoarelor din anii 1930 și 1940 în domeniul criptografiei și al comunicațiilor electrice. La Bell Labs s-a făcut pionierat în utilizarea matematicii în industrie. **Thornton Fry** (1892-1991) a fost directorul Departamentului de Matematică din 1924 și a fost convins de relevanța matematicii în noua eră industrială. În 1928 a publicat cartea *Probability and its Engineering Uses* și, în 1941, articolul „Industrial mathematics”, care poate fi citit în expoziția virtuală, menționată mai sus și în care a demonstrat importanța matematicii în comunicațiile electrice și în alte domenii ale industriei.

Contribuția lui Shannon s-a potrivit perfect spiritului pe care Fry l-a introdus la Bell Labs, aplicarea matematicii la problemele practice, fără a pierde din vedere ambiția intelectuală a matematicii pure. Cu alte cuvinte, nu a fost vorba despre găsirea unei soluții concrete unei probleme specifice, ci mai degrabă despre dezvoltarea unor cadre teoretice complete și generale care să permită o analiză cantitativă a unei familii întregi de probleme practice. Shannon a realizat acest lucru studiind mai întâi criptografia și apoi transmiterea mesajelor.

Claude Elwood Shannon s-a născut la 30 aprilie 1916, în Petoskey, SUA, un mic oraș turistic de pe malul Lacului Michigan. Fiul unei profesoare de limbi străine și al unui om de afaceri local, a studiat ingineria electrică și matematică la Universitatea din Michigan între 1932 și 1936, înainte de a se transfera la Massachusetts Institute of Technology (MIT), unde și-a finalizat studiile postuniversitare. Acolo și-a prezentat teza de masterat în 1937, despre care se spune adesea că este cea mai relevantă teză de masterat din secolul XX. Și aceasta nu este o exagerare. În lucrare a demonstrat cum se efectuează calcule logice, folosind circuite electrice care pot fi găsite în două stări: cu sau fără curent care curge prin ele. Aceasta este ideea de bază care a permis ulterior dezvoltarea calculatoarelor și a constituit primul contact al lui Shannon cu limbajul binar pe care va dezvolta ulterior teoria informației.

În 1940 și-a terminat doctoratul la MIT, cu o altă teză mai puțin cunoscută despre o formulare matematică a legilor geneticii. În același an a obținut o bursă pentru a-și continua cercetarea la Princeton, dar cu câteva luni mai târziu a decis să se alăture celor de la Bell Labs din Manhattan, unde a început să rezolve probleme de criptografie, folosind metode algebrice. În anii 1940, el a cercetat criptografia și comunicațiile, iar rezultatele cercetării sale s-au materializat într-o lucrare publicată în două articole în 1948, *O teorie matematică a comunicării*, în care a pus bazele teoriei informației. Sunt două articole de relevanță istorică deoarece nu numai că stabilesc cadrul teoretic

pentru studierea matematică a informațiilor și comunicațiilor, dar includ și două teoreme importante privind compresia datelor și capacitatea unui canal de a transmite mesaje.

Cu aceste articole, Shannon a câștigat imediat o recunoaștere enormă și binemeritată. Atât de mult, încât în 1949 a fost publicată lucrarea *The mathematical theory of communication*, o carte care a compilat cele două articole, precum și o expunere a teoriei mai accesibilă publicului larg și scrisă de **Warren Weaver** (1894-1978). Cartea lui Shannon și Weaver a atras interesul nu numai al inginerilor și al experților în comunicații, ci și al oamenilor de știință și al gânditorilor din toate domeniile.

Shannon a continuat să lucreze la teoria informației, deși rezultatele erau mai puțin substanțiale. Dezvoltarea în continuare a propriei sale teorii a fost realizată de matematicieni și de ingineri de la Bell Labs și din sute de centre de cercetare din întreaga lume. Laboratoarele Bell au fost suficient de mulțumite de el pentru a nu-i cere prea mult după munca sa din 1948. Șeful Departamentului de Matematică a mers atât de departe încât a spus: „Shannon și-a câștigat dreptul de a fi neproductiv”. I s-a permis să întârzie, să meargă pe coridoare pe un monociclu și să-și dedice timpul și bugetul oricărui proiect la care s-ar putea gândi. Cu astfel de condiții de muncă atât de puțin exigente, Shannon a abandonat aproape complet cercetările pe subiectele mai ortodoxe ale matematicii și ingineriei, dar a reușit să-și dezlănțuie imaginația, să proiecteze și să construiască mașini destul de curioase. Unele au avut un interes științific fără îndoială, precum un prototip capabil să joace șah, sau *Tezeu*, „șoarecele care rezolvă labirinturi”. Ambele pot fi considerate contribuții de pionierat în domeniul inteligenței artificiale. Un alt exemplu a fost aparatul de citire a minții. Acest artefact nu a fost un simplu amuzament, deși poate Shannon l-a conceput ca atare. Este una dintre primele mașini capabile să învețe. Ceea ce a învățat mașina lui Shannon au fost tiparele pe care ființele umane le urmează inconștient atunci când încearcă să imite soarta. Mașina este compusă dintr-un singur comutator care poate fi

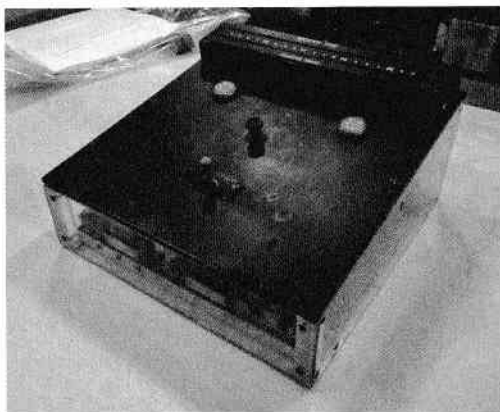
acționat spre dreapta sau spre stânga. Jucătorul rotește comutatorul și aparatul încearcă să prezică care va fi următoarea decizie. Astăzi ne putem juca cu diferite versiuni ale mașinii lui Shannon pe internet. De exemplu, doctorandul MIT Guy Satat a programat o replică a mașinii originale în acest link:

<http://web.media.mit.edu/~guysatat/MindReader/index.html>

Dacă vom încerca, vom descoperi că este destul de dificil să învingem. Noi, toți, prezentăm modele care, inconștient, se repetă în succesiunea de zerouri și unu pe care le generăm. Algoritmul pe care dispozitivul îl folosește pentru a descoperi tiparele adversarului său este precursorul unei discipline cu totul noi: învățarea automată (*machine learning*), extrem de utilă în inteligența artificială.

Alte mașini ale lui Shannon nu erau atât de utile în mod evident, dar aveau un fel de *poetică inginerescă*. THROBAC a fost un calculator care lucra cu cifre romane. *The ultimate machine* („mașina definitivă”) constă într-o cutie cu un singur comutator de alimentare. Când este acționat, dispozitivul pornește și apare o mână metalică din interiorul cutiei care duce comutatorul în poziția „oprit”. Singura acțiune a mașinii este să se oprească.

Figura 1. Mașina care citește mintea.



În 1956, Shannon a părăsit Bell Labs și s-a întors la MIT ca profesor, unde, în mare măsură, a menținut și a exercitat acel drept de a fi neproductiv. În ciuda acestui fapt, prelegerile sale au inspirat mulți studenți care au adus ulterior contribuții semnificative la teoria informației. De asemenea, el a investigat piețele financiare, jocurile de noroc și fizica jongleriei, a fost consultant în mai multe companii de tehnologie și, cum nu putea fi altfel, a continuat să construiască artefacte fără utilitate aparentă, cum ar fi automatele de jonglerie sau mașinile care au rezolvat cubul lui Rubik. De asemenea, și-a păstrat personalitatea inconfundabilă și simțul umorului până în anii următori. În 1981, într-o scrisoare adresată editorului de la *Scientific American*, cerându-i scuze pentru că nu a putut să livreze la timp un articol despre jonglerie, a scris: „Sunt un poet mai bun decât om de știință”. Mesajul a inclus o scurtă poezie despre cubul lui Rubik, intitulată „A rubric on Rubik cubics”. La scurt timp, au apărut primele simptome ale Alzheimerului, o boală care l-a ținut departe de viața academică în ultimii săi ani de viață. La 4 februarie 2001, la o reședință din apropierea casei sale din Massachusetts, Claude Shannon a încetat din viață la vârsta de 84 de ani.

Atitudinea intelectuală a lui Shannon a fost un amestec unic de două pasiuni aparent opuse: dorința de a juca, de a proiecta și de a construi mașini reale cu funcții specifice și, pe de altă parte, o pasiune pentru abstractizarea matematicii și pentru capacitatea enormă pe care ni le oferă când putem găsi formalismul potrivit pentru a reprezenta problemele practice în termeni matematici. Această atitudine s-a dovedit a fi cea mai potrivită pentru a face față provocărilor noii ere a informației.

Shannon a reușit, cu doar două lucrări, să construiască bazele întregii teorii a informației. Au existat multe evoluții ulterioare, dar contribuția sa rămâne nucleul fundamental al teoriei. Concepțiile informaționale și aplicarea lor la compresia datelor (prima teoremă a lui Shannon sau teorema de codare a sursei), precum și capacitatea de transmisie a canalelor de comunicare (a doua teoremă a

lui Shannon sau teorema de codare a canalului) au apărut deja în articolele din 1948 și le vom dedica primele trei capitole ale cărții. Discutând compresia datelor în capitolul al doilea, vom putea, de asemenea, să aprofundăm aspecte care s-au dezvoltat după acele articole de pionierat, precum codurile Huffman, algoritmul de compresie Lempel-Ziv sau complexitatea algoritmică. În cel de-al treilea capitol vom aprofunda teoria comunicării și vom explica în detaliu a doua teoremă a lui Shannon și aplicațiile sale în telefonie mobilă și în alte sisteme de comunicații. Teoria informației a lui Shannon a avut un impact considerabil asupra mai multor probleme fundamentale din fizică, cum ar fi axa timpului, iar în ultimele trei decenii s-a descoperit că sistemele cuantice pot procesa informațiile mult mai eficient decât cele clasice, ceea ce a dat naștere la o disciplină cu totul nouă: informațiile cuantice. Vom dedica ultimul capitol al cărții acestei relații neașteptate dintre fizică și informație; vom completa astfel toată gama de aplicații și de concepte, inspirate din teoria care s-a născut în 1948 din mâinile lui Claude Shannon.

Capitolul 1

Bazele teoriei informației

Sens, informație și aleatoritate

Primul contact cu teoria informației tinde să genereze o confuzie considerabilă și o oarecare dezamăgire. Ne așteptăm ca teoria să se ocupe matematic de mesajele pe care le schimbăm și chiar de propria gândire și de modul în care este formulată în limbaj.

Totuși, primul lucru pe care îl înveți este ceva surprinzător: faptul că un text cu litere aleatorii, fără niciun model, conține mult mai multe informații decât *Don Quijote* sau decât *Enciclopedia Britanică*. Motivul este foarte simplu. Textul lui *Don Quijote* este plin de redundanțe. Dacă apare litera *q*, știm că următoarea va fi *u*. La *Dulcin* va urma *ea*, cel mai probabil. În schimb, textul aleatoriu este imprevizibil. Nu există nicio redundanță sau posibilitate de a deduce care este următoarea literă din cele precedente. În teoria matematică a informației, importantă este imprevizibilitatea, deoarece nu se preocupă de semnificația mesajelor sau a textelor, ci, mai degrabă, încearcă să cuantifice informațiile pe care le pot deține. Din acest motiv, informațiile din textul aleatoriu sunt mai multe decât cele prezente într-un paragraf al lui *Don Quijote* de aceeași lungime.

De fapt, prima mare contribuție a lui Shannon este tocmai aceasta: analiza comunicării, schimbul de mesaje, luând în considerare doar proprietățile lor statistice și nu semnificația lor. Shannon însuși afirmă acest lucru în mod explicit în primul său articol din 1948: „Aspectele semantice ale comunicării sunt irelevante pentru problema inginerescă. Aspectul semnificativ este că mesajul trimis este unul selectat dintr-un set de mesaje posibile”.¹

Dacă ne ocupăm doar de proprietățile statistice ale mesajelor, atunci putem să-l considerăm pe expeditorul său ca o sursă aleatorie, cu alte cuvinte, ca și cum ar fi luat în mod aleatoriu simboluri ale unui alfabet (litere, cifre, zerouri și unu etc.) în funcție de anumite probabilități: „Ne putem imagina o sursă discretă care generează mesajul, simbol cu simbol. El alege fiecare simbol în funcție de anumite probabilități care depind, în general, de alegerile precedente și de simbolul care trebuie ales”.²

Prin urmare, teoria informației nu se referă la mesaje individuale, ci la surse sau la generatoare de simboluri aleatorii cu anumite proprietăți statistice. Odată ce reducem sursele de informație la generatori de simboluri aleatorii, următorul pas este de a cuantifica informațiile pe care aceste simboluri le conțin și care vor depinde doar de proprietățile statistice, menționate anterior.

Ideea lui Shannon este de a defini informațiile ca numărul mediu de zerouri și de unu (adică biți) necesari pentru a descrie mesajele generate de sursă sau, mai general, pentru a descrie rezultatul oricărui eveniment aleatoriu. Dar sarcina de a descrie ceva folosind zerouri și unu este foarte asemănătoare cu ceea ce se întâmplă în anumite jocuri de ghicit ale copiilor. Astfel încât, cea mai bună modalitate de a înțelege definiția *informației* lui Shannon este să analizăm în detaliu unul dintre aceste jocuri.

¹ [The] semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages.

² We can think of a discrete source as generating the message, symbol by symbol. It chooses successive symbols according to certain probabilities depending, in general, on preceding choices as well as the particular symbols in question.

Întrebări și răspunsuri

Vă amintiți jocul în care trebuie să ghiciți un personaj familiar folosind întrebări cărora li se poate răspunde doar cu da sau cu nu? Să presupunem că ghicim personajul ascuns, de exemplu, Marie Curie, prin cinci întrebări: Este bărbat? Nu. Este europeană? Da. Este artistă? Nu. Este om de știință? Da. S-a născut în Polonia? Da. Aceasta înseamnă că putem codifica figura Marie Curie folosind 5 cifre binare, 01011, unde fiecare 0 indică un răspuns negativ și fiecare 1, un răspuns pozitiv. Această idee poate fi extinsă la orice personaj, obiect, eveniment sau număr ascuns: dacă elaborăm un protocol de întrebări de tip da/nu pentru a rezolva enigma, răspunsurile vor forma un șir de zerouri și unu, adică de biți, care codifică toate obiectele ascunse posibile (care ar trebui să formeze un set finit).

Rețineți că fiecare răspuns pe care îl primim în joc restricționează numărul de candidați posibili. Când răspunsul este afirmativ sau negativ cu aceeași probabilitate, adică $1/2$, întrebarea împarte posibilitățile candidaților în două grupuri aproximativ egale, masculin și feminin, de exemplu, iar răspunsul elimină unul dintre cele două grupuri. Aceste tipuri de întrebări sunt tocmai cele care ne oferă cele mai multe informații, în medie, despre personaj. Cu alte cuvinte, întrebarea optimă este cea al cărei răspuns are aceleași probabilități, adică $1/2$, de a fi da sau de a fi nu. Pentru a ne convinge de acest lucru, în loc să ghicim un personaj, să ne imaginăm că trebuie să ghicim un număr între 0 și 7. Să vedem ce se întâmplă dacă punem întrebări în care probabilitatea celor două răspunsuri este $1/2$. Am putea începe, de exemplu, întrebând dacă numărul ascuns este mai mare sau egal cu 4 (adică dacă este 4, 5, 6 sau 7). Răspunsul, oricare ar fi fost, ne-ar permite să respingem patru numere și să le păstrăm pe celelalte patru drept candidați. Următoarea întrebare ar trebui să respingă două numere. De exemplu, dacă ni s-a spus că numărul este mai mare sau egal cu 4, a doua întrebare ar putea fi: