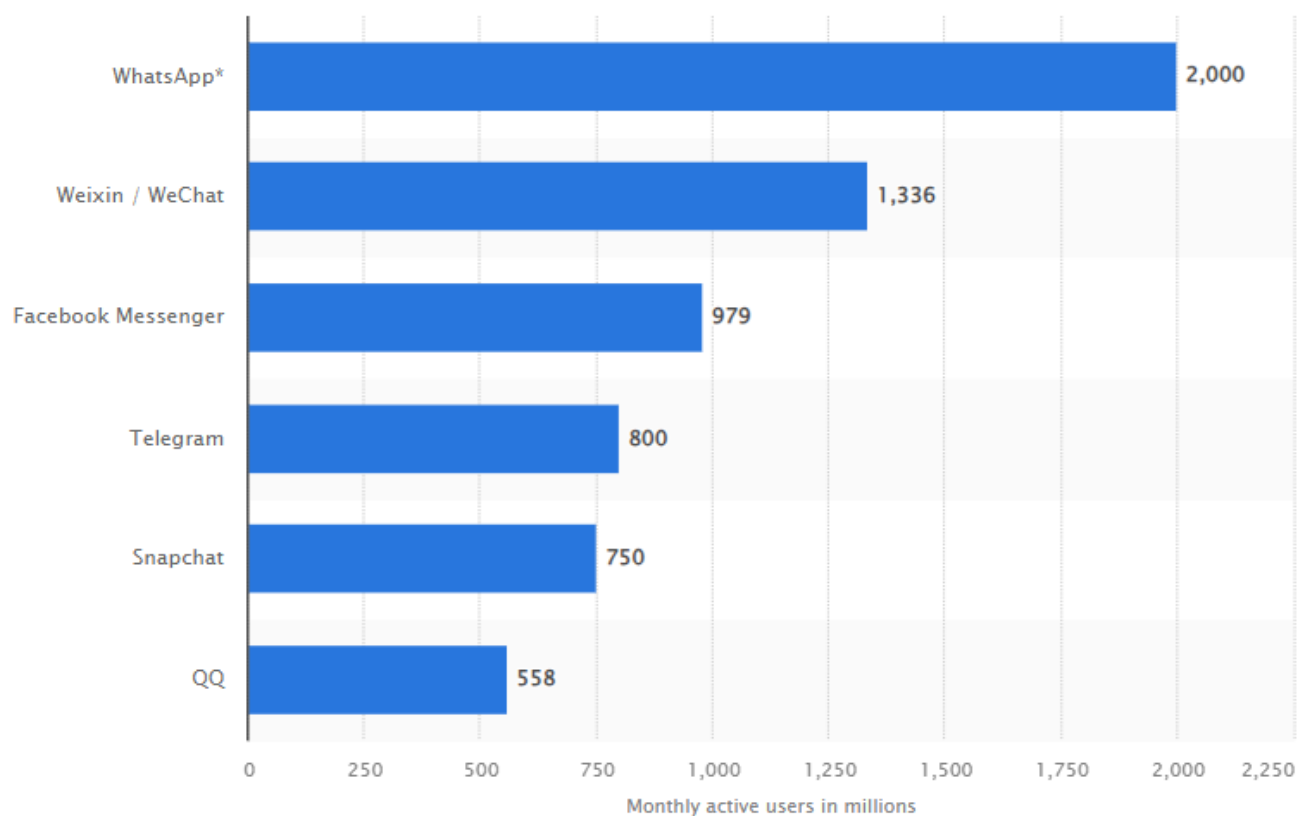


Introducere in data mining

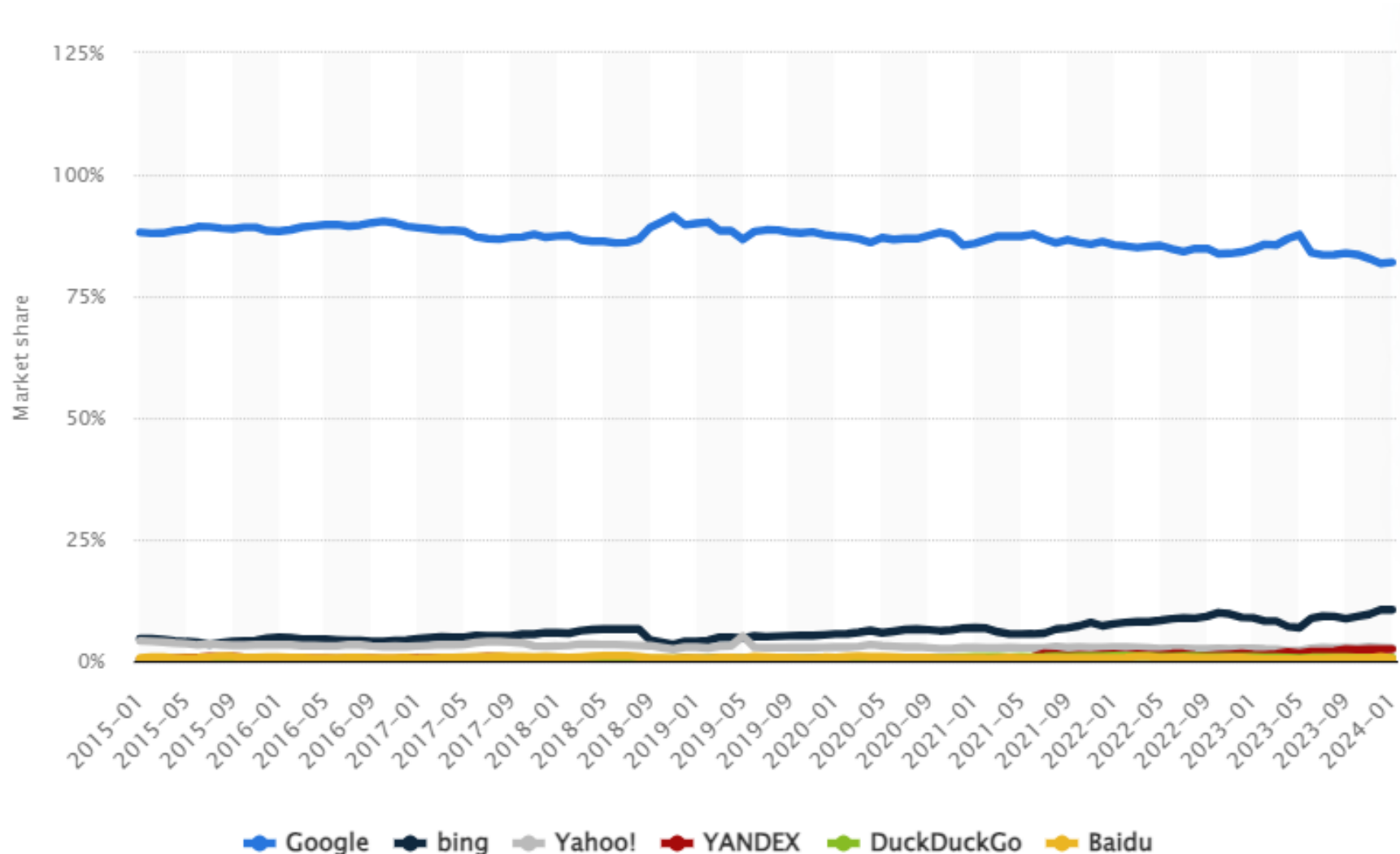
- Odata cu explozia tehnologica, cu introducerea calculatoarelor in toate domeniile de activitate, cu extinderea conexiunilor high-speed la internet, cu dezvoltarea domeniului IoT, cu dezvoltarea spatiilor de stocare a informatiilor, s-a creat un urias bagaj de informatii si o retea de transfer de volume uriase de date.
- Cu cat mai multe dispozitive electronice sunt inventate si interconectate, cu atat volumul datelor este mai mare.
- Ca rezultat, știința si industria trebuie să facă față provocării de a lucra cu seturi de date de mari dimensiuni.

Cele mai populare aplicații globale de mesagerie mobilă din ianuarie 2024, pe baza numărului de utilizatori activi lunar (în milioane)



<https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/>

Cota de piață a principalelor motoare de căutare desktop la nivel mondial din ianuarie 2015 până în ianuarie 2024

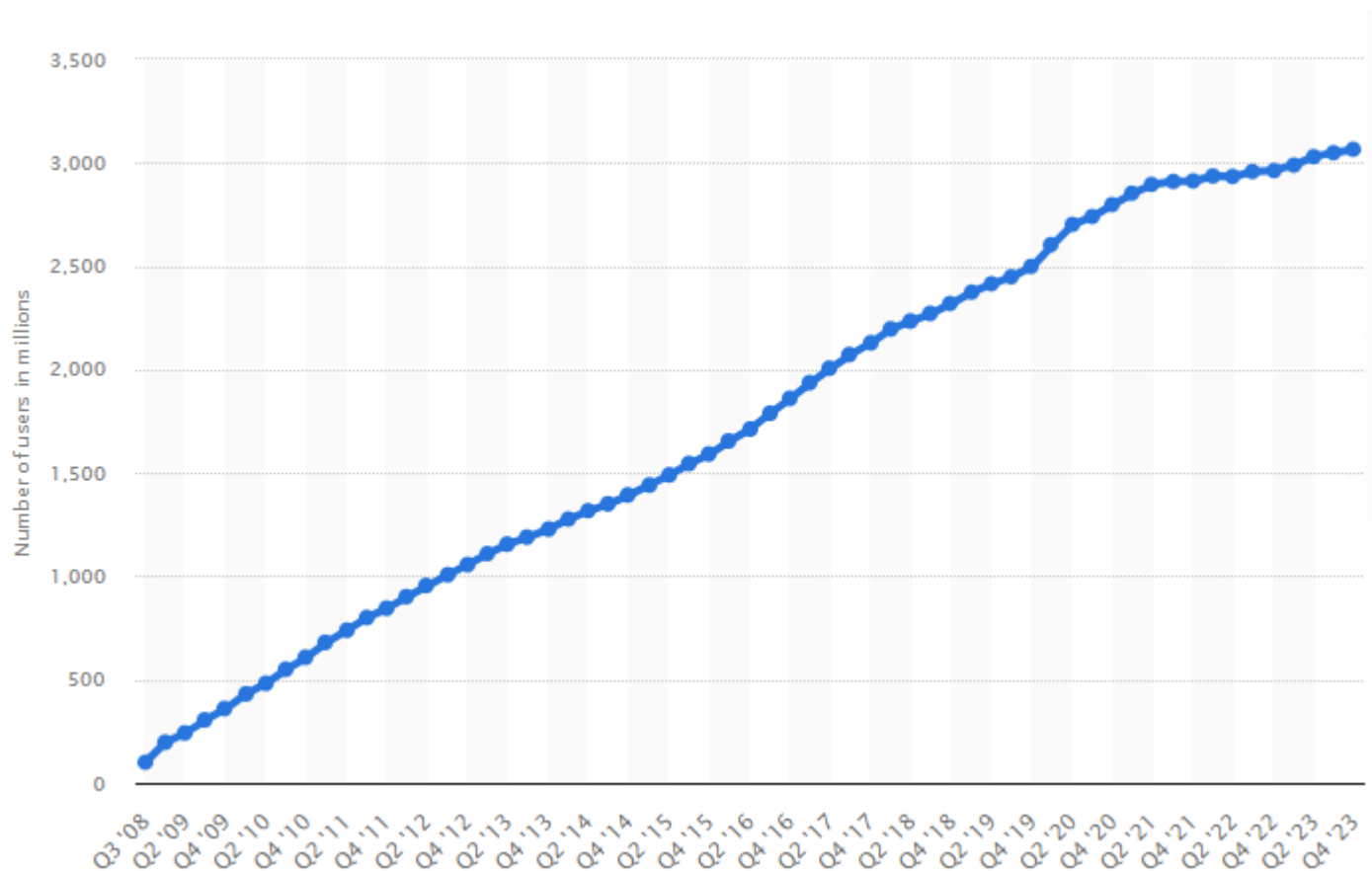


<https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/>

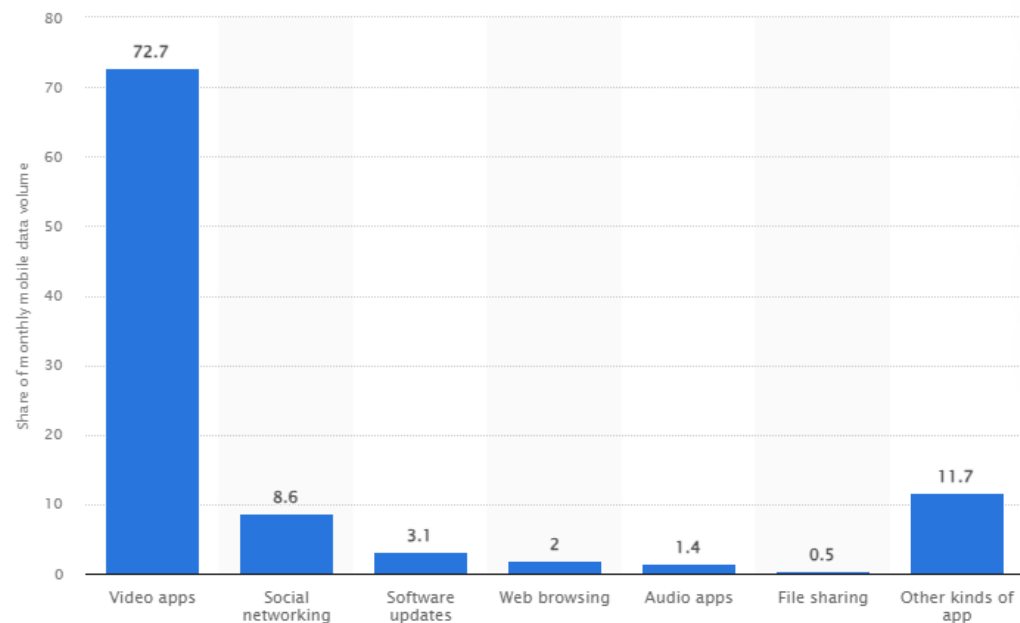
<https://gs.statcounter.com/search-engine-market-share>



Numărul de utilizatori Facebook activi lunar la nivel mondial începând cu trimestrul 4 2008 (în milioane)

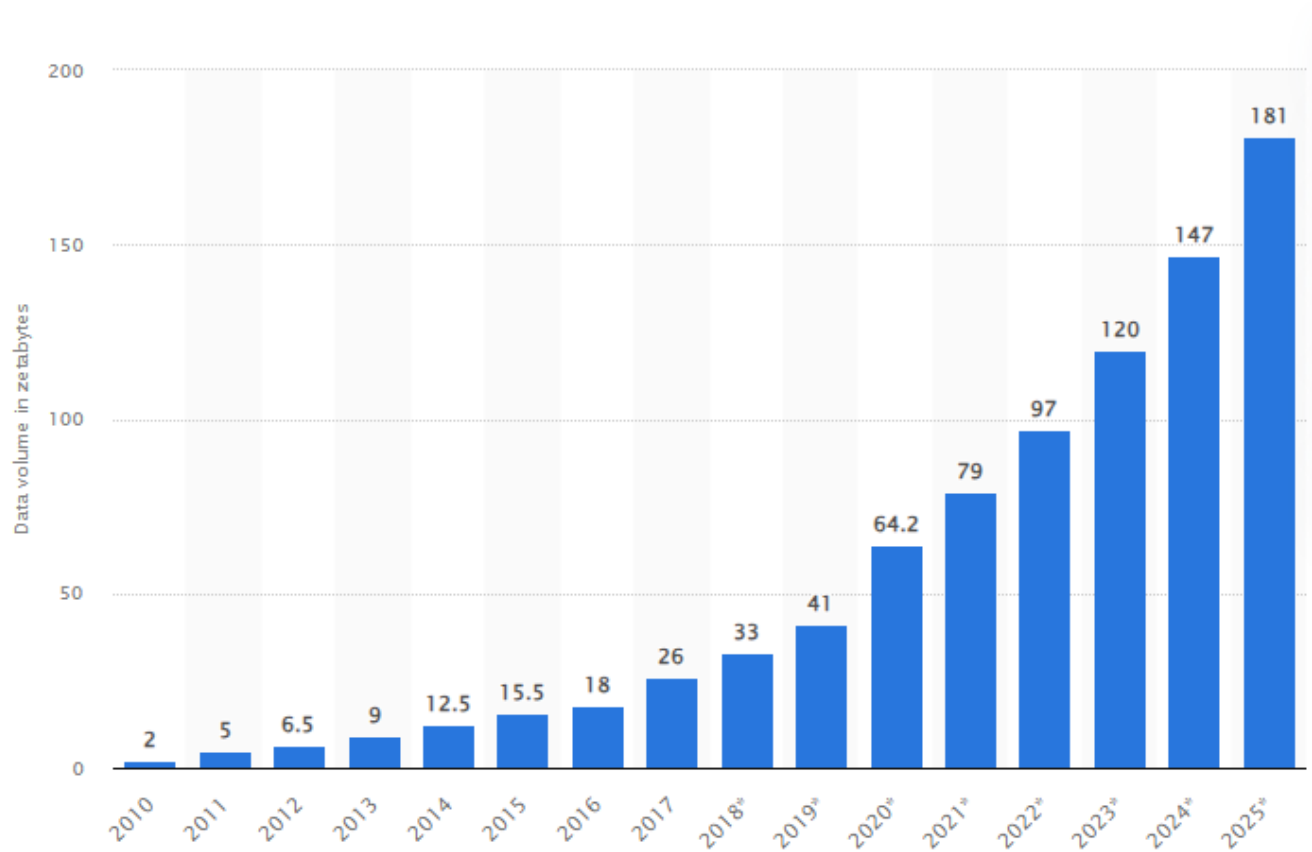


Distribuția volumului lunar global de date pentru aplicațiile mobile începând cu ianuarie 2024



- <https://www.statista.com/statistics/383715/global-mobile-data-traffic-share/>

Volumul de date/informații create, capturate, copiate și consumate la nivel mondial din 2010 până în 2020, cu previziuni din 2021 până în 2025 (în zetabytes)

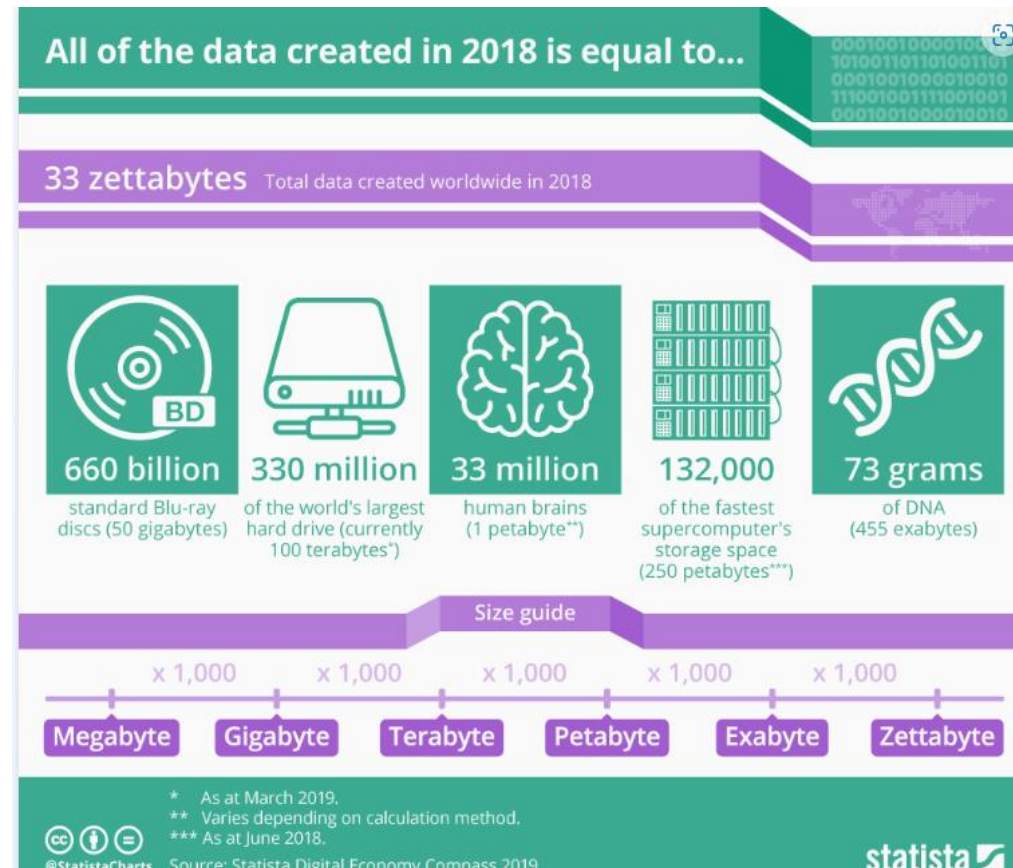


1 ZB = 1000 exabyte (EB) = 2^{10} EB

= 1.000.000 petabyte (PB) = 2^{20} PB

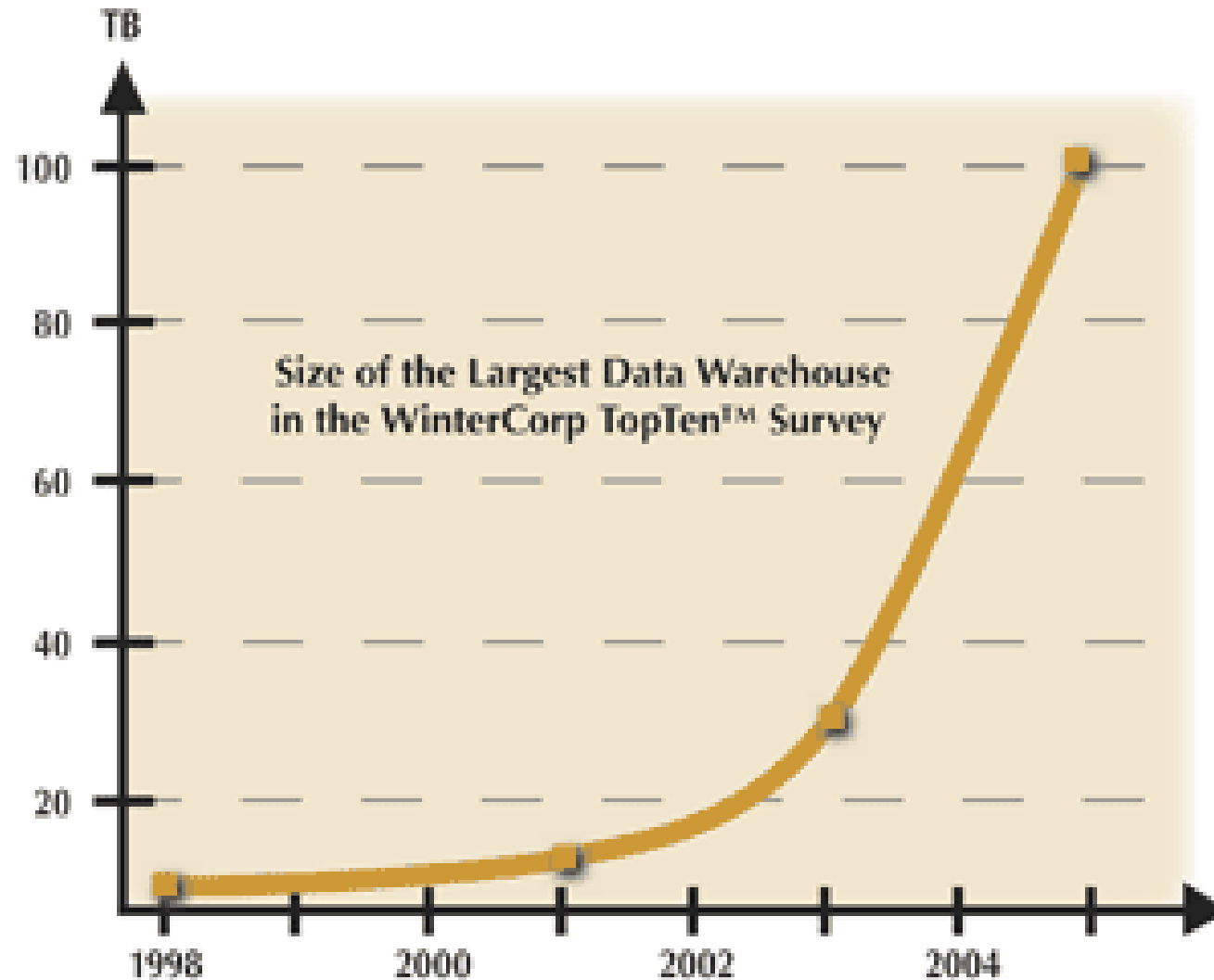
= 1.000.000.000 terabyte (TB) = 2^{30} TB

= 1.000.000.000.000 gigabyte (GB) = 2^{40} (GB)



[Chart: All of the data created in 2018 is equal to... | Statista](#)

Dimensiunea bazelor de date



- 1 TB = 1024GB

Date, date, date...

- În timp ce acești "munți" de date sunt ușor de produs în zilele noastre, este dificil de "sapat", de minerit pentru a obține informații valoroase din aceștia.
- Datele colectate sunt din ce în ce mai complexe, dinamice și greu de analizat. Un volum tot mai mare de date este colectat, depășind cu mult capacitatea dispozitivelor și metodelor software de a le captura, organiza, procesa în timp relativ scurt.

Ce este data mining?

- Data mining este analiza automata a datelor, in general a bazelor de date de dimensiuni mari, cu scopul de a descoperi tendinte, sabloane, tipare netriviale, necunoscute anterior, uneori neasteptate, in date si care ar putea oferi informatii utile.
- Data mining ofera algoritmi si tehnicile necesare interpretarii bazelor de date de dimensiuni mari.

Introducere in data mining

- Dificultatea analizei bazelor de date f mari
 - Datele provin din diverse surse si sunt in formate foarte variate: text, video, imagini, sunete
 - Tehnicile de colectare a datelor sunt variate.
 - Procesul de extragere a informatiilor trebuie sa fie eficient si uneori aproape in real-time.



Tehnici data mining

Tehnici specifice domeniului data mining folosite pentru analiza automata a datelor.

Data warehousing

- Organizarea datelor intr-o maniera consistenta si folositoare = data warehousing
- Magaziile de date (*data warehouses*) si modul in care sunt organizate, modul in care lucreaza cu date, cu informatiile incomplete, de exemplu, sunt foarte importante pentru data mining.
- Magaziile de date ofera "memorie" informatiilor, in timp ce data mining ofera "inteligenta".

- Pentru ca aplicarea metodologiei data mining sa aiba succes sunt necesare:
 - o magazie de date bine organizata si integrata
 - o intelegere foarte buna a fenomenului/experimentului caruia metodologia este aplicata.

METODOLOGIA DATA MINING

Modelarea = procesul de creare a unui model pentru o situatie in care se cunoaste raspunsul si de aplicare a acelu model intr-o situatie in care nu se cunoaste raspunsul.

Exemplu

- Sunteți directorul de marketing al unei companii de telecomunicații și vreți să racolați clienți noi.
- Puteți foarte simplu să:
 - vă faceți cunoscuți prin pliante/ads pe care să le distribuiți prin poșta/email sau în magazinele specializate, pe rețelele de socializare etc. sau
 - să folosiți experiența de afaceri înmagazinată deja în baza dvs de date pentru a crea un model.


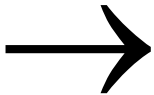
Exemplu

- Informatii despre toti clientii firmei:
 - varsta, sex, adresa
 - modul de folosire a serviciului (daca isi platesc factura la timp, daca folosesc serviciul des, adica daca vorbesc mult la telefon si cat profit aduc firmei).
- Folosind datele despre clientii actuali, puteti crea un model pe care il puteti testa folosind tot baza de date pe care o aveti despre clientii actuali.

Modelare

- Pentru a face asta, in procesul de modelare data mining, anumite date se pun deoparte, deci se foloseste doar o parte a bazei de date, parte care trebuie sa fie suficient de cuprinzatoare, totusi, se creaza modelul si se testeaza pe datele care au fost puse deoparte, pentru a-l valida.
- Abia apoi se aplica pentru luarea de noi decizii privind promotia ce se doreste.
- Folosind acest model, directorul de marketing poate decide sa se adreseze promotional numai unei parti a populatiei.

Exemplu

- Pentru a face asta, in procesul de modelare data mining, anumite date se pun deoparte
- (2/3 din baza de date)  multimea de training
- se creaza modelul
- se testeaza pe datele care au fost puse deoparte
- (1/3 din baza de date)  multimea de testare
- pentru a-l valida.
- Abia apoi se aplica pentru luarea de noi decizii privind promotia ce se doreste.

Istoric

- **1989** – workshop-uri organizate de ACM ([Association for Computing Machinery](#)) in **cadrul conferintelor** Special Interest Group (SIG) on Knowledge Discovery in Databases
- Fondatori:
 - **Usama Fayyad**
 - **Gregory Piatetsky-Shapiro**
 - **Rakesh Agrawal**

Istoric

Usama Fayyad - autor al cartilor de referinta in domeniul data mining

- *Fayyad, Usama (1996). [Advances in Knowledge Discovery and Data Mining](#) (1 ed.) AAAI Press. Reeditata in 2010.*
- *Fayyad, Usama (2002). [Information Visualization in Data Mining and Knowledge Discovery](#) (1 ed, Morgan Kaufmann Publishers. Reeditata in 2010*
- Yahoo,
- acum: Executive Director for the Institute of Experiential Artificial Intelligence., *Northeastern University, US*
- Editor al revistelor de specialitate
 - Data Mining and Knowledge Discovery – cel mai important jurnal din domeniu
 - SIGKDD Explorations – newsletter
 - Legends of Data & AI - podcast

Istoric

- **Gregory Piatetsky-Shapiro** – Presedinte al [Kdnuggets](#) (pana in 2021), website pt Analytics, Big Data, Data Science, Data Mining si Machine Learning.

U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, eds., [Advances in Knowledge Discovery in Databases](#), AAAI/MIT Press 1996.

Nr. 1 in [LinkedIn Top Voices 2018: Data Science & Analytics](#) si inclus in [Top Artificial Intelligence Influencers To Follow in 2019](#)

Istoric

- **Rakesh Agrawal** (IBM Research, initial)
 - A lucrat la mai multe produse IBM, cel mai important: Intelligent Miner
 - Articolul [Fast algorithms for mining association rules](#) in colaborare cu [Ramakrishnan Srikant](#) publicat in Proceedings conferintei [VLDB](#) in 1994 si care descrie algoritmul Apriori, este unul din articolele cele mai influente din domeniul bazelor de date.

Big data

- Definitie: “*Big data reprezinta date de o mare **varietate** care sunt generate in **volume** din ce in ce mai mari și cu o **viteza** mai mare decat niciodata.*” (Gartner)

(the three V's variety-volume-velocity)

- Aceste seturi de date sunt atât de voluminoase încât software-ul tradițional de prelucrare a datelor pur și simplu nu le poate gestiona.
- *Challenges and Opportunities with Big Data*
 - Lucrare realizata de mai multi cercetatori (Computing Community Consortium)
 - <http://cra.org/ccc/docs/init/bigdata/whitepaper.pdf>2012
-

Data science vs. Knowledge Discovery in Databases

Data science

- **Data science** este explorarea si analiza cantitativa a tuturor tipurilor de date, structurate sau nestructurate, pentru intelegerea acestora, extragerea de informatii si formularea de rezultate care pot fi puse in aplicare.

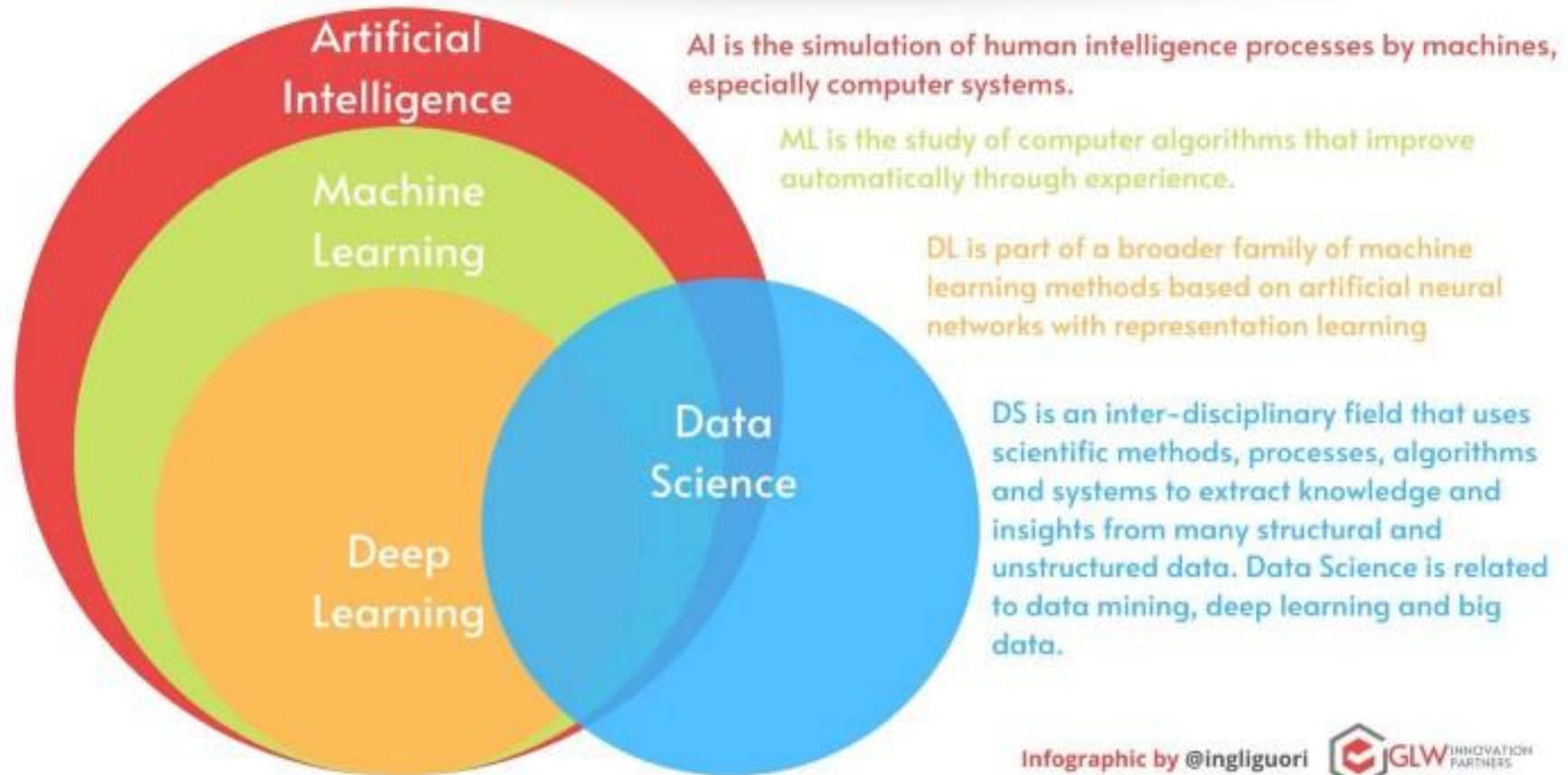
Knowledge Discovery in Databases

- Knowledge Discovery in Databases este procesul de extragere de cunostinte din bazele de date, data mining reprezentand doar un pas in acest proces.
 - Fayyad, Piatetsky-Shapiro, Smyth, *From Data Mining to Knowledge Discovery: An Overview*, AAAI Press / The MIT Press, 1996

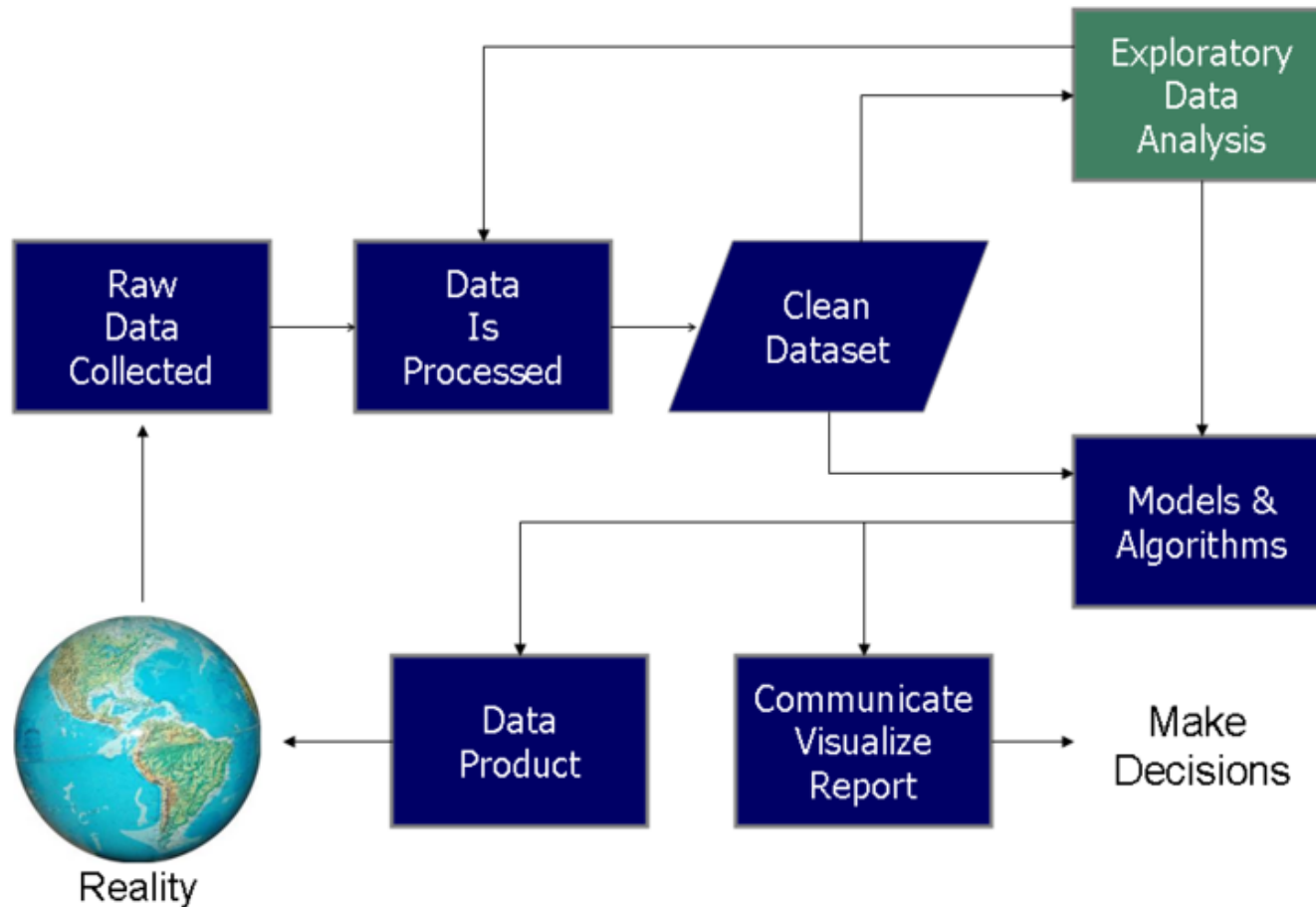
- Data mining își propune să înțeleagă și să descopere cunoștințe noi, nedescoperite anterior în date.
- Inteligența artificială (IA) este preocupată de construirea mașinilor inteligente cu scopul de a crea sisteme automate care se comportă la fel ca oamenii.
- Învățarea automata (ML) este o parte a inteligenței artificiale și își propune să dezvolte algoritmi bazati pe multimi de training și pe experiență.
- Învățarea profundă (Deep learning) este o parte a inteligenței artificiale , în care datele sunt transmise prin aplicarea a numeroase transformări neliniare pentru a genera un output.

RELATIONSHIP BETWEEN

ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, DEEP LEARNING AND DATA SCIENCE



Data Science Process



[Sursa:Farcaster/](#) [English Wikipedia](#)

Big data vs. data science

- <http://www.kdnuggets.com/2015/07/data-science-big-data-different-beasts.html>
- “Collecting Does Not Mean Discovering”

Domeniile de aplicabilitate

- Domeniile de activitate in care se pot aplica tehnicile data mining:
 - in stiinta: astronomie, medicina
 - in domeniul afacerilor (comercial): managementul relatiei cu clientii (CRM –customer relationship management), comertul on-line, telefonie, sport si entertainment, marketing, investitii
 - Web: motoare de cautare, text si web mining
 - Securitatea cibernetica (Cybersecurity)

EXEMPLE DE APLICARE A METODOLOGIEI DATA MINING

- Churn (customer attrition) – renuntarea unui client la serviciile oferite de bussines-ul respectiv
- (un client al companiei X renunta la serviciile de telefonie ale acesteia – paraseste compania- nu stim de ce si nici daca)
- Sa presupunem urmatoarea situatie:
 - O companie de telefonie mobila observa ca anual 25-30% din clientii sai renunta la serviciile companiei. Ce isi propune compania?
 - Sa estimeze asa numita valoare a clientului (prin valoare intelegand, de exemplu, client bun (foloseste des serviciile, plateste la timp, factura este mare etc.) sau in caz contrar, client cu valoare mica)
 - si in functie de valoarea estimata sa pregateasca o oferta sau o modalitate de a-l pastra pe client, sa faca ceva pt ca acel client sa nu renunte la servicii.

- **credit risc**

Sa consideram urmatoarea situatie:

O persoana aplica pentru un credit bancar. Ce risc isi asuma banca? Ar trebui sa ii aprobe sau nu creditul? Bancile folosesc pentru aceasta o serie de programe pentru estimarea riscului si luarea unei decizii.

- **comertul on-line**

- Situatie: Cineva vrea sa cumpere o carte. Atunci i se ofera sa cumpere si alte carti, pe care probabil le va cumpara, mai exact, carti cu cea mai mare probabilitate de cumparare de catre acel client. Adica oferte personalizate.

- **Medicina**

- Printr-un pateneriat, IBM si Mayo Clinic, una din cele mai renumite clinici din US, si din lume de altfel, cu spitale in trei state din US , lucreaza din 2004 la un proiect de data mining cu scopul de a gasi metode de tratament individualizate pentru fiecare pacient.
- Ei investigeaza o baza de date pe care au creat-o si pus-o la punct, baza de date ce cuprinde pe langa datele personale ale fiecarui pacient, rezultate de laborator, rezultatele electrocardiogramelor, ale radiografiilor.
- Proiectul isi propune sa gaseasca sabloane, tendinte legate de modul in care pacientii, in functie de varsta, bolile pe care le-au avut, precum si alti factori, raspund la diverse tratamente.
- Aceasta va permite doctorilor sa aleaga tratamentul cel mai bun pentru un pacient, tratamentul care a avut cel mai mare succes pentru pacienti cu aceleasi similaritati ca si pacientul in cauza.
- Se spera ca pana in 2014, medicii clinicii Mayo sa poata folosi aplicatiile acestui proiect. In septembrie 2011 anuntau ca sunt aproape gata cu implementarea tehnicilor data mining.
- Ideea este de fapt de a aplica cunostintele despre mai multi pacienti, pentru ca un pacient sa beneficieze de cel mai bun tratament. Se propune de asemenea imbunatatirea bazei de date cu informatii genetice ale pacientilor.

- IBM Watson este un supercomputer care combina inteligența artificială cu softuri de analiză a datelor pentru a reproduce capacitatea funcțională umană de a răspunde la întrebări.
- Pentru aceasta Watson accesează 90 de servere cu o capacitate de stocare de peste 200 milioane de pagini, în total, pe care le procesează la o rată de 80 de teraflops .
- Parteneriat IBM - Mayo Clinic - Doctorii să folosească capacitatea lui Watson de a crea protocoale de tratament personalizat pentru pacienți cu boli cronice (incurabile).
- Watson Health și Pfizer Inc. au anunțat o colaborare ce va utiliza IBM Watson pentru descoperirea de noi medicamente în domeniul immuno-oncologiei
- <https://www-03.ibm.com/press/us/en/pressrelease/46768.wss>
- Martie 2021 - IBM a lansat un parteneriat de 10 ani cu gigantul medical academic Cleveland Clinic pentru a aplica capacitățile sale de calcul cuantic, cloud și inteligență artificială în dezvoltarea medicamentelor și cercetarea agenților patogeni.
- <https://www.healthcaredive.com/news/ibm-cleveland-clinic-launch-10-year-quantum-computing-deal/597546/>

Medicina

- Google Flu Trends
- nov 2008 –in US, si alte 3 tari
oct 2009-alte 16 tari, multe din Europa
actualizat in 2009, 2013, 2014

<https://www.google.org/flutrends/about/>

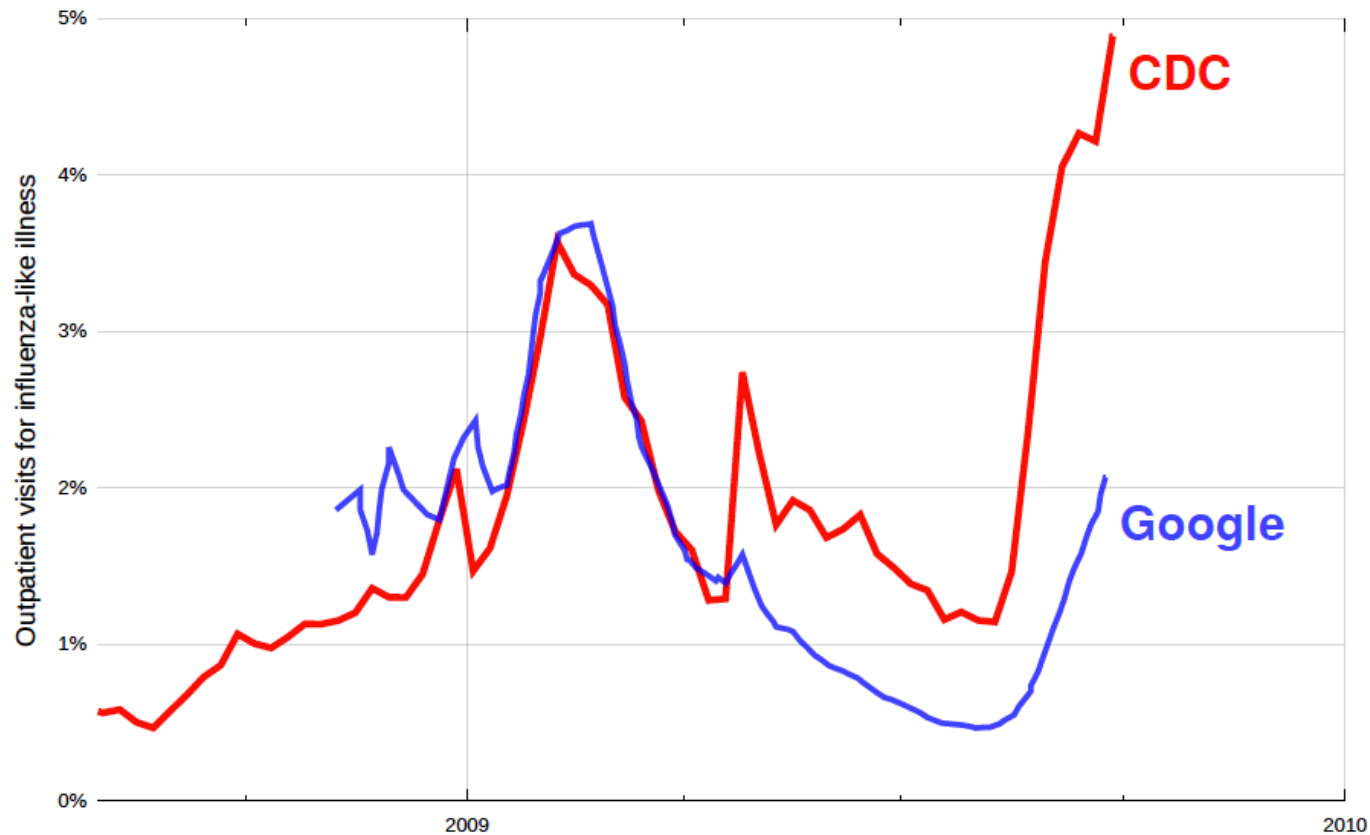
https://en.wikipedia.org/wiki/Google_Flu_Trends

- Descoperind o relatie stransa intre nr de oameni cauta pe google termeni legati de gripa (flu) si cati oameni au efectiv simptomele acestei gripe, au reusit sa urmareasca modul de distributie a gripei sezoniere in US, mai repede decat Centre for Disease Control and Prevention (CDC). (cu o intarziere de 24 de ore, fata de 1 saptamana cat este necesar pt CDC)
- Ei au reusit sa estimeze nivelul de gripa aproape in real-time, dupa numarul de cautari ale anumitor termeni.
- Google doar a rulat termenii de căutare pe care utilizatorii ii introduceau in motorul de cautare. Algoritmii rulati au dus la obtinerea de rezultate.
- <http://www.theneweconomy.com/strategy/big-data-is-not-without-its-problems>

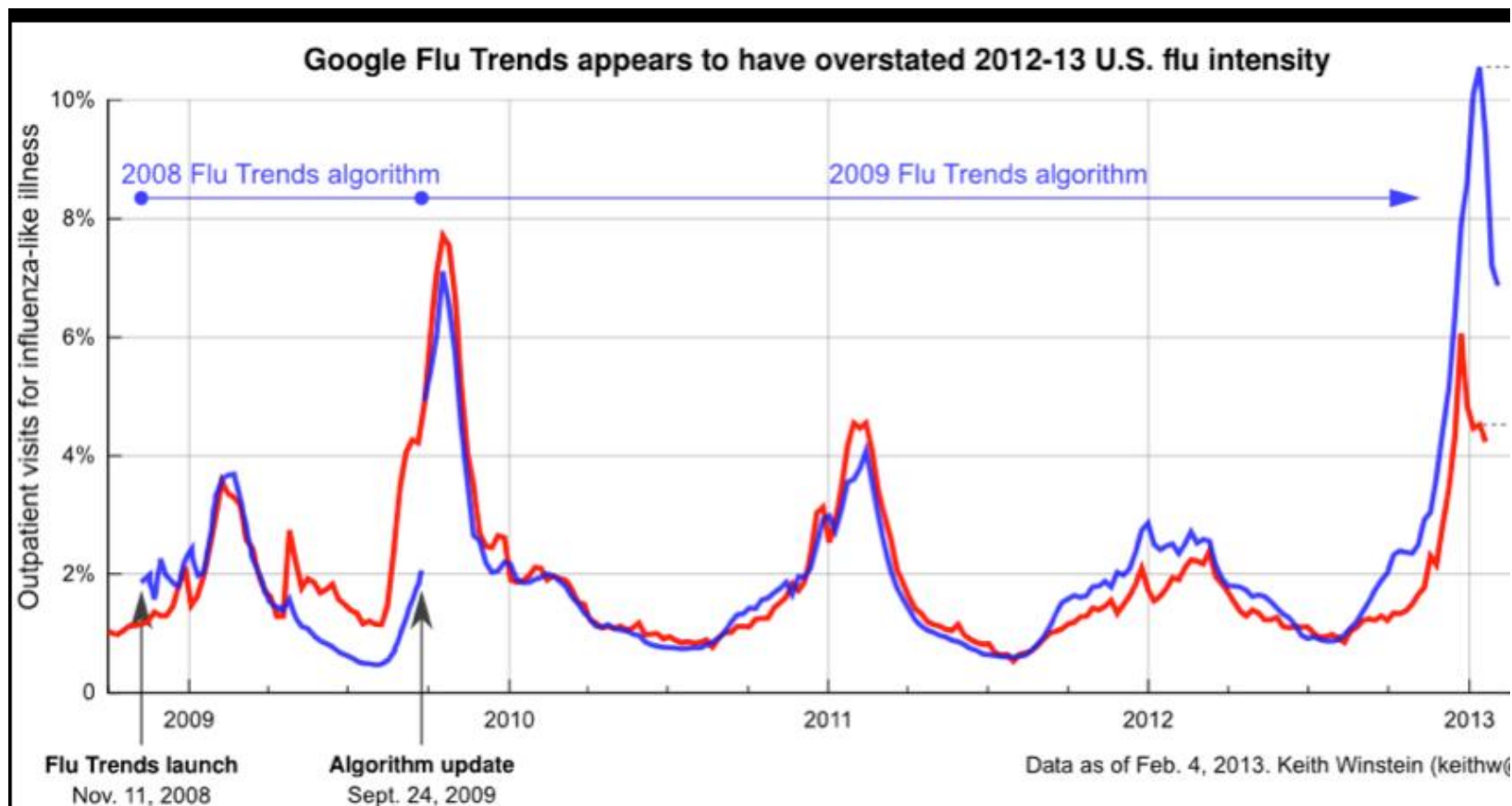
- La inceput, modelul propus de Google parea sa fie f. bun, rezultatele lor fiind comparabile cu cele ale CDC([Centers for Disease Control and Prevention](#))
- Din pacate a subestimat epidemia din 2009 a virusului H1N1

Sursa:

<http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html>



Sursa: <http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html>



Modelul a starnit o alarma falsa estimand ca nr de cazuri de gripa din anul 2012-2013 ar fi dublu decat a fost in realitate.

- Insa cercetatorii incearca sa gaseasca motivul pentru care modelul propus de Google a esuat si sa gaseasca noi solutii de estimare a cazurilor de gripa sezoniera.
- **Samuel Kou**, un profesor de la Harvard mentioneaza faptul ca in timp, oamenii isi schimba stilul, modul in care cauta informatii despre acelasi lucru in motoarele de cautare.
- Acesta, impreuna cu o echipa, au creat un model de data mining care, testat pe datele din 2009-2015 a dat rezultate f. bune.
- Modelul propus de ei combina datele propuse de Google cu informatii despre gripele sezoniere si informatii obtinute de la CDC.
- <https://news.harvard.edu/gazette/story/2015/11/on-top-of-the-flu/>
- *“Advances in using Internet searches to track dengue”*, 2017
- “Using electronic health records and Internet search information for accurate influenza forecasting”, 2017
- 2019-prezentare la Universitatea din Chicago
- **Unmasking the Actual COVID-19 Case Count**, **Author:** Kou, Samuel C; Yang, Shihao, Clinical Infectious Diseases, Oxford University Press, 2020-05-15

- **detectarea unor infractiuni,**
 - folosirea unor carti de credit in mod fraudulos,
 - efectuarea unor convorbiri telefonice frauduloase.
 - Detectarea atacurilor cibernetice
- Aproape toate tranzactiile efectuate cu cartile de credit sunt scanate cu ajutorul unor algoritmi specifici, pentru a identifica tranzactiile suspicioase.
- Marile companii de telefonie folosesc software specializate pentru identificarea convorbirilor frauduloase.

Campania electorala a lui Obama 2012

- S-a bazat pe un model analitic ce foloseste tehnici de data mining construit de o echipa condusa de Daniel Wagner, 29 de ani, pentru a prezice care sunt oameni care il vor vota pe presedintele Obama.
- Pentru construirea modelului au contactat prin telefon mii de votanti pentru a-i identifica pe cei care l-ar vota.
- Apoi au analizat ce au aceste persoane in comun.
- Baza de date a inclus peste 80 de attribute, inclusiv varsta, sexul, modul in care a votat in anii anteriori, daca are casa, unde locuieste, ce fel de casa/apartament are, la ce reviste este abonat.

- Acest model i-a ajutat pe membrii campaniei lui Obama ca, in mod eficient, sa recruteze voluntari, sa trimita email-uri sau scrisori de sustinere, sa participe la activitati de strangere de bani.
- Modelul a identificat in fiecare oras persoanele care vor vota cu democratii. Ei au reusit sa ii atraga pe cei indecisi, in statele care nu erau hotarate, precum Virginia, Florida, and Ohio.
- Modelul a estimat ca va fi reales. Rezultatele estimate au fost foarte apropiate de realitate.
- Campania lui Obama a schimbat modul in care campaniile prezidentiale se vor desfasura in viitor.
- In campania din 2016 si Clinton si Trump au folosit modele de analiza a datelor. ([*Gregory Piatetsky, Kdnuggets: Trump, Failure of Prediction, and Lessons for Data Scientists*](#))
- <http://www.kdnuggets.com/2016/11/trump-shows-limits-prediction.html>

Legatura dintre data mining si alte domenii

- Unii spun ca data mining nu este decat un nume dat statisticii.
- Este adevarat ca multe din tehnicile folosite in data mining provin din alte domenii: statistica, informatica, inteligenta artificiala. Cu toate acestea, exista diferente intre aceste domenii si data mining.
- Statistica ofera un suport teoretic pentru studiul evenimentelor intamplatoare si metode pentru testarea ipotezelor, dar nu studiaza preprocesarea datelor sau vizualizarea rezultatelor, ce fac parte din data mining.
- Inteligenta artificiala are o alta metoda de aprofundare, mai euristica si se concentreaza pe imbunatatirea performantei agentilor de invatare. Data mining se concentreaza pe intregul proces de descoperire de cunostinte, de la organizarea datelor si eliminarea celor incomplete, invatare si cunoastere prin descoperire si pana la vizualizarea rezultatelor.

Metode data mining

- **Supervizate** (supervised) - se propune explicarea sau divizarea pe categorii a unui atribut al bazei de date. Exista un atribut tinta.
- **Nesupervizate** (unsupervised) - se propune descoperirea de sabloane, relatii sau similaritati intre grupuri de inregistrari fara ca sa se foloseasca o colectie de clase predefinite sau un camp target.

Operatii specifice data mining

- Clasificare
- Estimare (Regresie)
- Asociere
- Clusterizare
- Descriere si profil

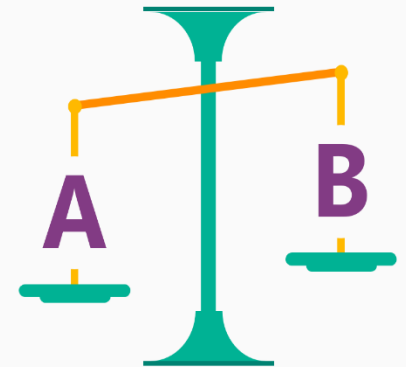
Clasificarea

- A clasifica inseamna a examina trasaturile si caracteristicile unui obiect si a-l repartiza unui set de clase predefinite.
- Clasificarea se evidentiaza printr-o caracterizare foarte bine definita a claselor si o multime de training ce consta in exemple preclasificate. Clasificarea consta in construirea unui model care sa poata fi aplicat unor date neclasificate inca tocmai pentru a putea fi clasificate.
- Tehnicile data mining folosite pentru clasificare sunt clasificarea Bayesiană, arborii de decizie, tehnicile de tipul cel mai apropiat vecin.

- Sursa: <https://docs.microsoft.com/en-us/azure/machine-learning/studio/data-science-for-beginners-the-5-questions-data-science-answers>

Is this A or B?

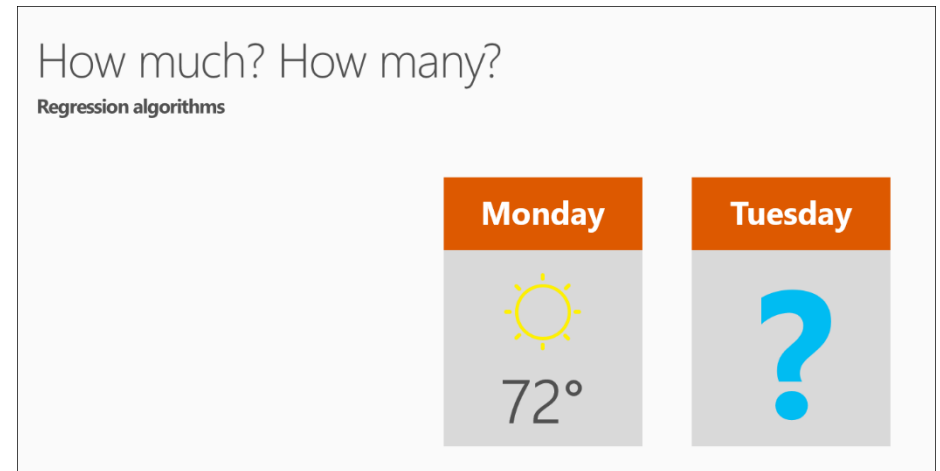
Classification algorithms



ESTIMARE (REGRESIE)

- In timp ce clasificarea lucreaza cu rezultate de tip discret, estimarea lucreaza cu rezultate cu valori continue.
- Date anumite date de intrare, estimarea determina o valoare necunoscuta inca pentru o variabila de tip continuu.
- Estimarea se foloseste foarte des pentru a clasifica inregistrările.
- De exemplu, se construiește un model care, bazat pe mult mai multe date, sa poata estima pentru fiecare client care ar fi probabilitatea de a raspunde promotiei si apoi clientii sa fie ordonati descrescator in functie de aceasta probabilitate, iar primii clienti, cu cele mai mari probabilitati, sa fie cei alesi pentru a primi oferta publicitara.
- Tehnicile data mining folosite pentru estimare sunt: regresie si rețele neurale.

Sursa: <https://docs.microsoft.com/en-us/azure/machine-learning/studio/data-science-for-beginners-the-5-questions-data-science-answers>



Asociere

- Asocierea este operatia de a determina ce lucruri pot fi grupate impreuna.
- De exemplu, determinarea produselor care sunt cumparate impreuna intr-un supermarket.
- Asocierea este o modalitate de a genera anumite reguli de forma "daca se intampla acest lucru atunci se va intampla si acest lucru (cu probabilitatea x).
- De exemplu,
 - "Persoanele care cumpara lapte vor cumpara si cereale cu probabilitatea p"
 - "Persoanele care cumpara o planta vor cumpara si pamant pentru flori in 60% din cazuri iar, ambele produse au fost cumparate impreuna in 5% din cazuri"
 - "Persoanele care cumpara o papusa Barbie vor cumpara si o ciocolata cu probabilitatea 60%".
-



Sursa:

<https://twitter.com/analyticbridge/status/658850748164280320/photo/1>

Clusterizare

- Clusterizarea este operatia de segmentare a unei multimi eterogene intr-un numar de subgrupuri mai omogene numite clustere.
- Clustering este de obicei o operatie preludiu la o alta operatie de data mining.
- De exemplu, clustering este prima operatie care se executa in dezvoltarea de strategii de segmentare a pietei.
- In loc sa ne punem intrebarea "Care este tipul de promotie la care clientii raspund cel mai bine?", intai impartim clientii in clustere cu obiceiuri de a cumpara produse asemanatoare si apoi adresam intrebarea "Care este tipul de promotie la care clientii dintr-un anumit cluster raspund cel mai bine?"

How is this organized?

Clustering Algorithms



Sursa: <https://docs.microsoft.com/en-us/azure/machine-learning/studio/data-science-for-beginners-the-5-questions-data-science-answers>

Descriere si profil

- Uneori scopul metodologiei data mining este de a descrie:
 - ce se intampla,
 - care sunt tendintele in baza de date pentru o mai buna intelegere a persoanelor, produselor si proceselor care au dus la producerea datelor din baza de date.
- Printre tehnicile folosite se numara arborii de decizie, regulile de asociere si clustering.