SYSTEMS LABORATORY

Curs DS01

Introducere în Știința Datelor (Data Science).
Măsurare

PR network crunch

Structura cursului

- Introducere: De ce DS?
- Scopul cursului
- Măsurarea și colectarea datelor
- Erori de măsurare
  - Validitatea
  - Fidelitatea
- Erori de modelare
  - Corelații iluzorii
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erori de eșantionare
- Tehnologiile datelor: programare vs. Învățare automată
- Procesul învățării automate
- Predicție și cunoaștere: analiza datelor
  - Măsurare și reducerea dimensionalității
  - Clasificare (cluster)
  - Cauzalitate și regresie
  - Predicții și serii de timp
- Predicția ca premisă pentru acțiune
- Predicția ca produs economic

Cuprins

#1.1 Introducere

- La ce folosește Știința Datelor?
- Scopul cursului

#1.2 Măsurarea și colectarea datelor

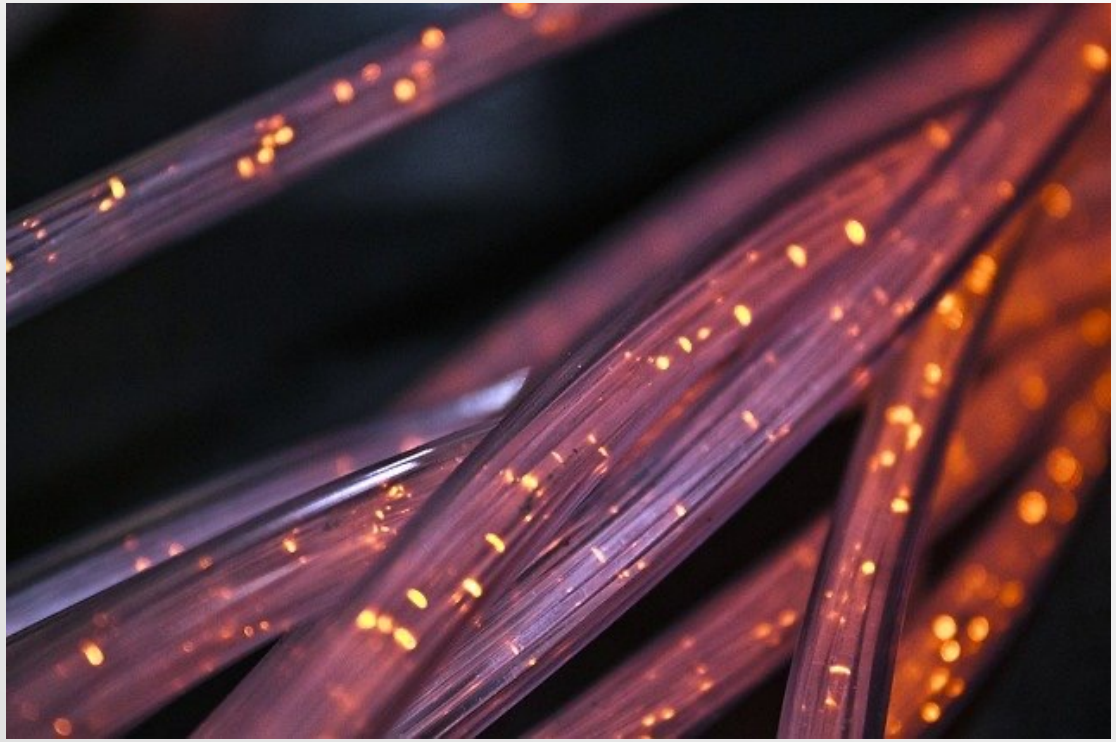
#1.3 Erori de măsurare

#1.1 Intro: De ce Data Science?

Data Science și Proiectarea Rețelelor: Analiză trafic, Detectare atacuri (DDOS), Schimbare dinamică a calității serviciului (QOS), Clasificare Trafic, Schimbare funcționalitate SDN(rețele definite prin software)/NFV (funcții virtuale de rețea)

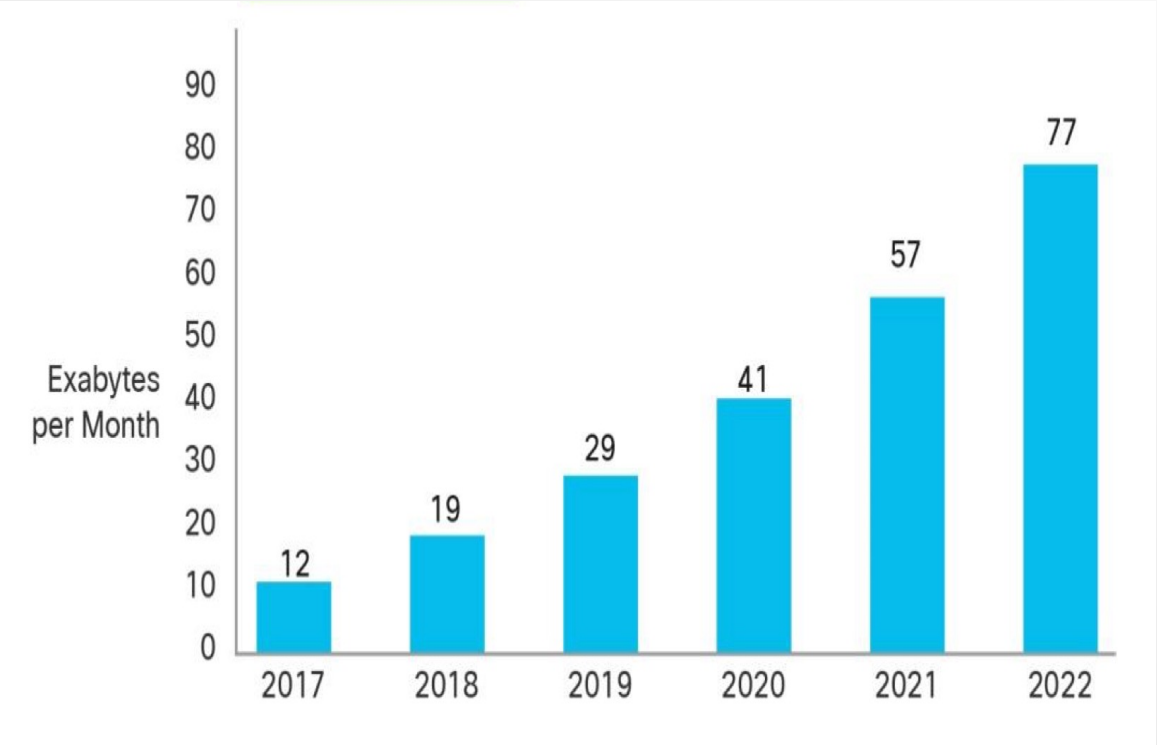
Data Science & PR

- Analiză trafic [Lakhina et al]
- Detectare atacuri (DDOS) [Gu et al.]
- Schimbare dinamică a calității serviciului (QOS) [Dekeris et al., Kuang et al]
- Clasificare Trafic [Erman et al.]
- Schimbare funcționalitate SDN(rețele definite prin software)/NFV (funcții virtuale de rețea) [Le et al.]



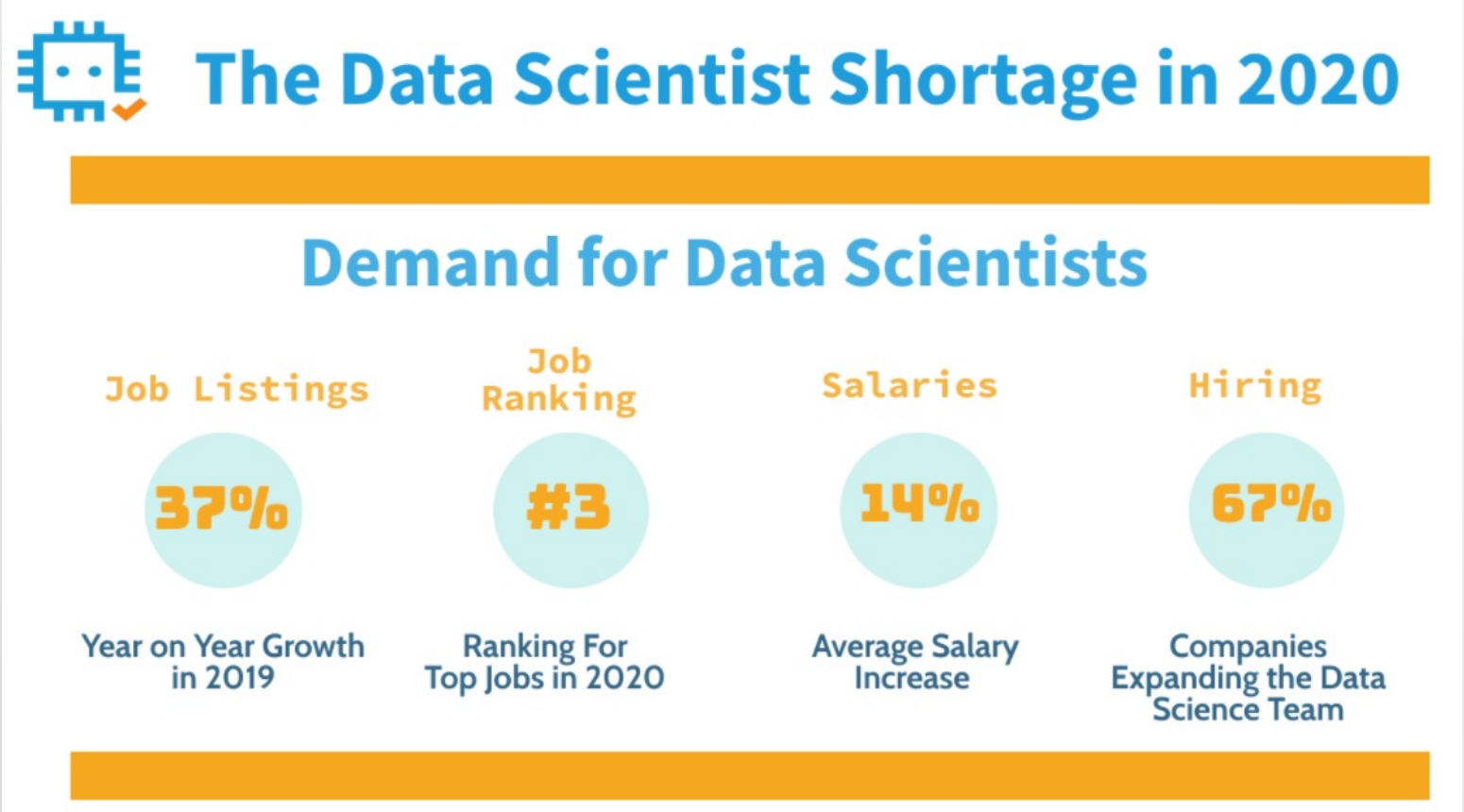
Creșterea Traficului

- Creștere de 70% pe an doar pentru traficul de dispozitive mobile [Cisco]

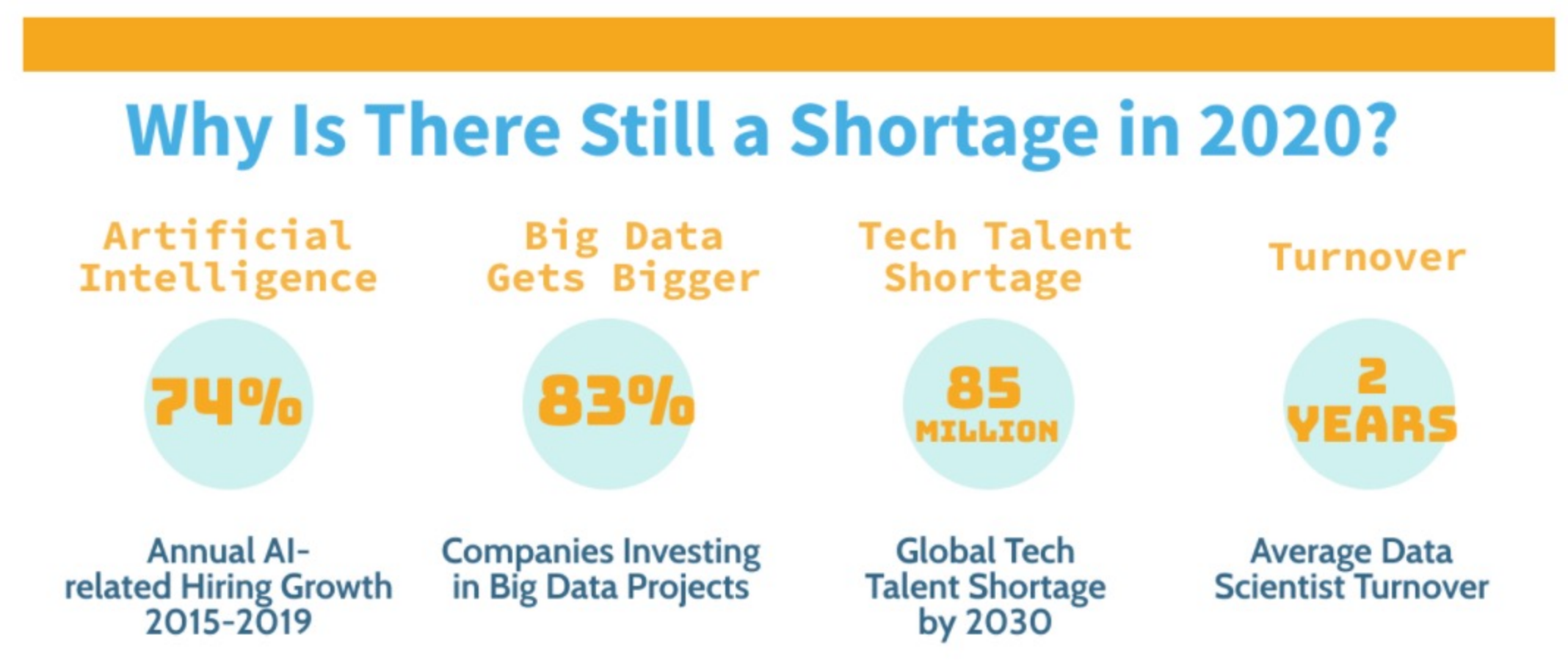


Importanța Științei Datelor - date

- Cererea este și va fi mult mai mare decât oferta
  - Bigger Data
  - AI



[Quanthub]


Why Is There Still a Shortage in 2020?
Artificial Intelligence
74%
Annual AI-related Hiring Growth 2015-2019
Big Data Gets Bigger
83%
Companies Investing in Big Data Projects
Tech Talent Shortage
85 MILLION
Global Tech Talent Shortage by 2030
Turnover
2 YEARS
Average Data Scientist Turnover

[Quanthub]

Difficulty Finding Security & Data Science Skillsets

Q: In which technology-related areas do you anticipate your organization will have the most difficulty in finding appropriate skillsets?

[Quanthub]

De la analiza datelor la știința datelor

The six Vs of big data

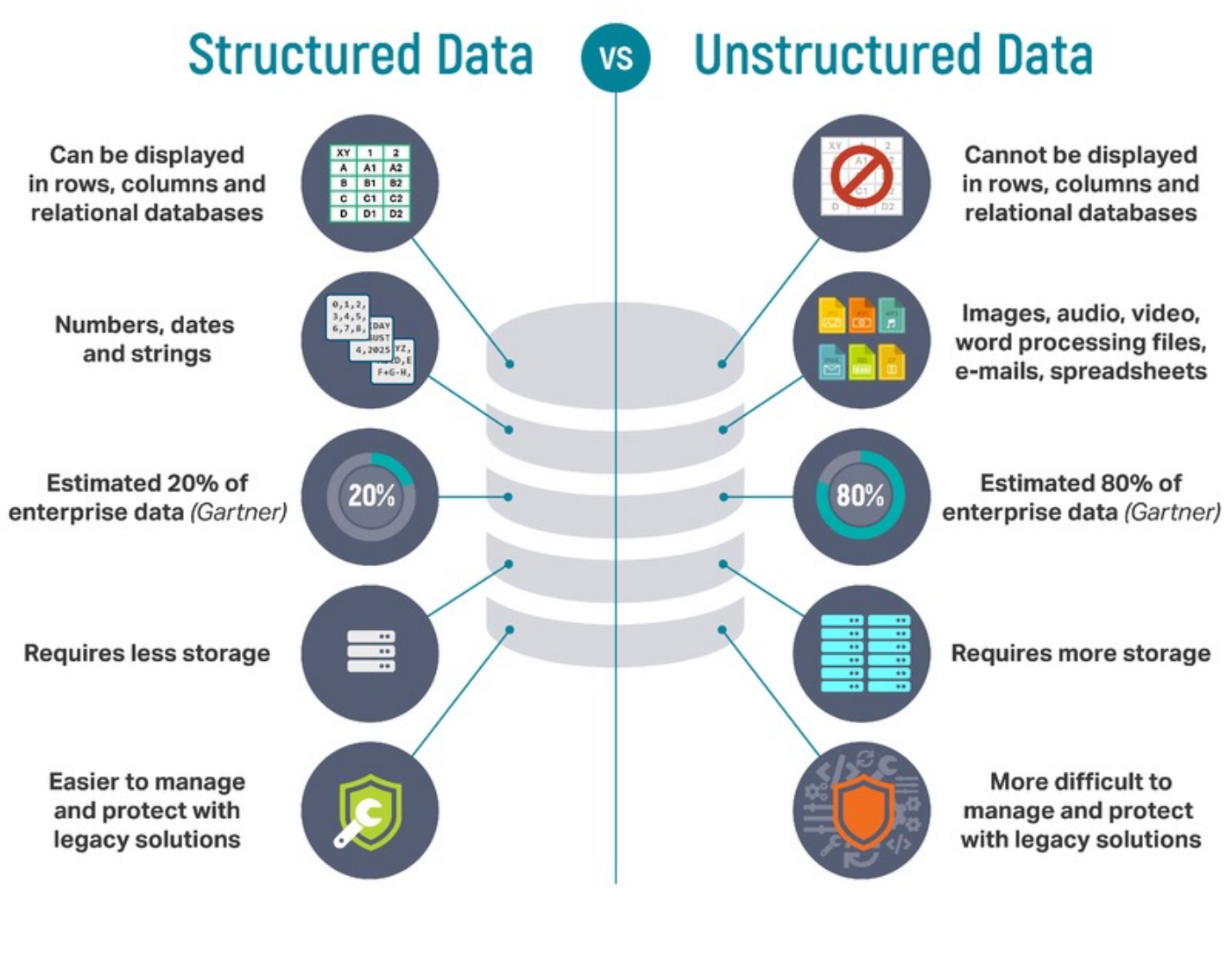
Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.

ICONS: ALEXBNDZ/ADOBE STOCK

©2018 TECHTARGET. ALL RIGHTS RESERVED TechTarget

Date structurate vs. nestructurate



[Chiang, 2018]

Business intelligence vs. știința datelor

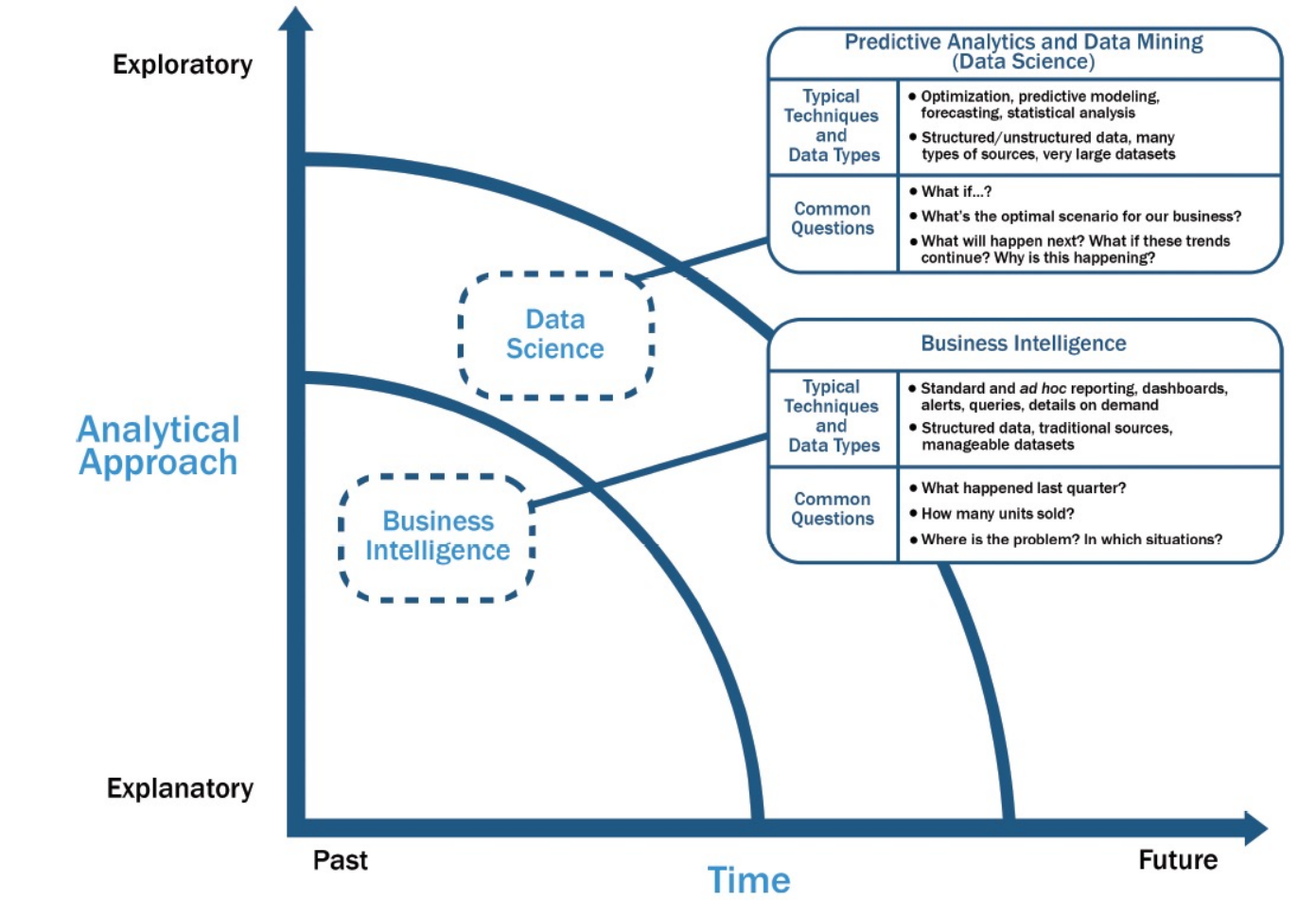


FIGURE 1-8 *Comparing BI with Data Science*

[EMC Education services]

Utilitatea în business a științei datelor

TABLE 1-2 *Business Drivers for Advanced Analytics*

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX)

[EMC Education services]

Ecosistemul științei datelor

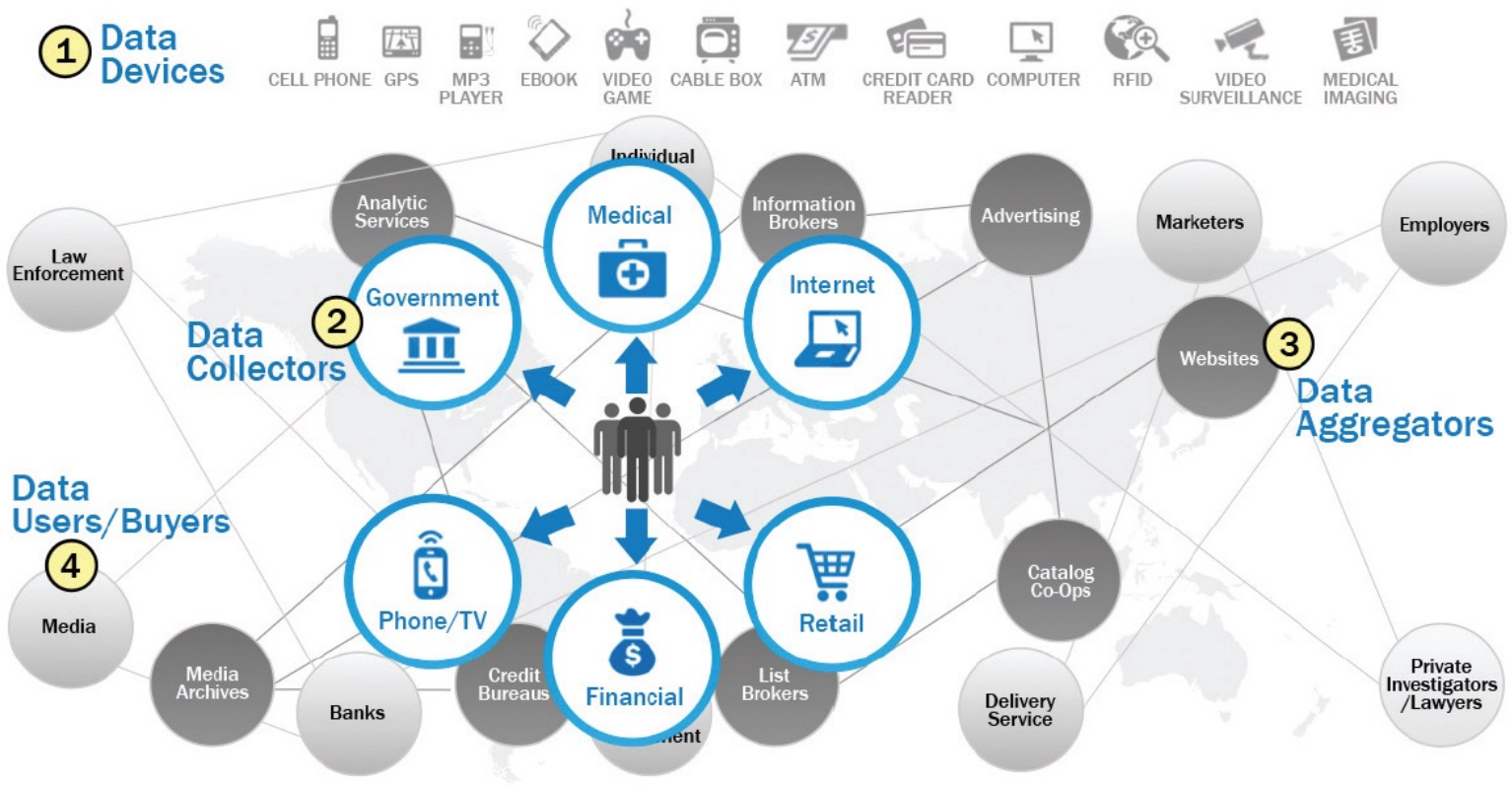


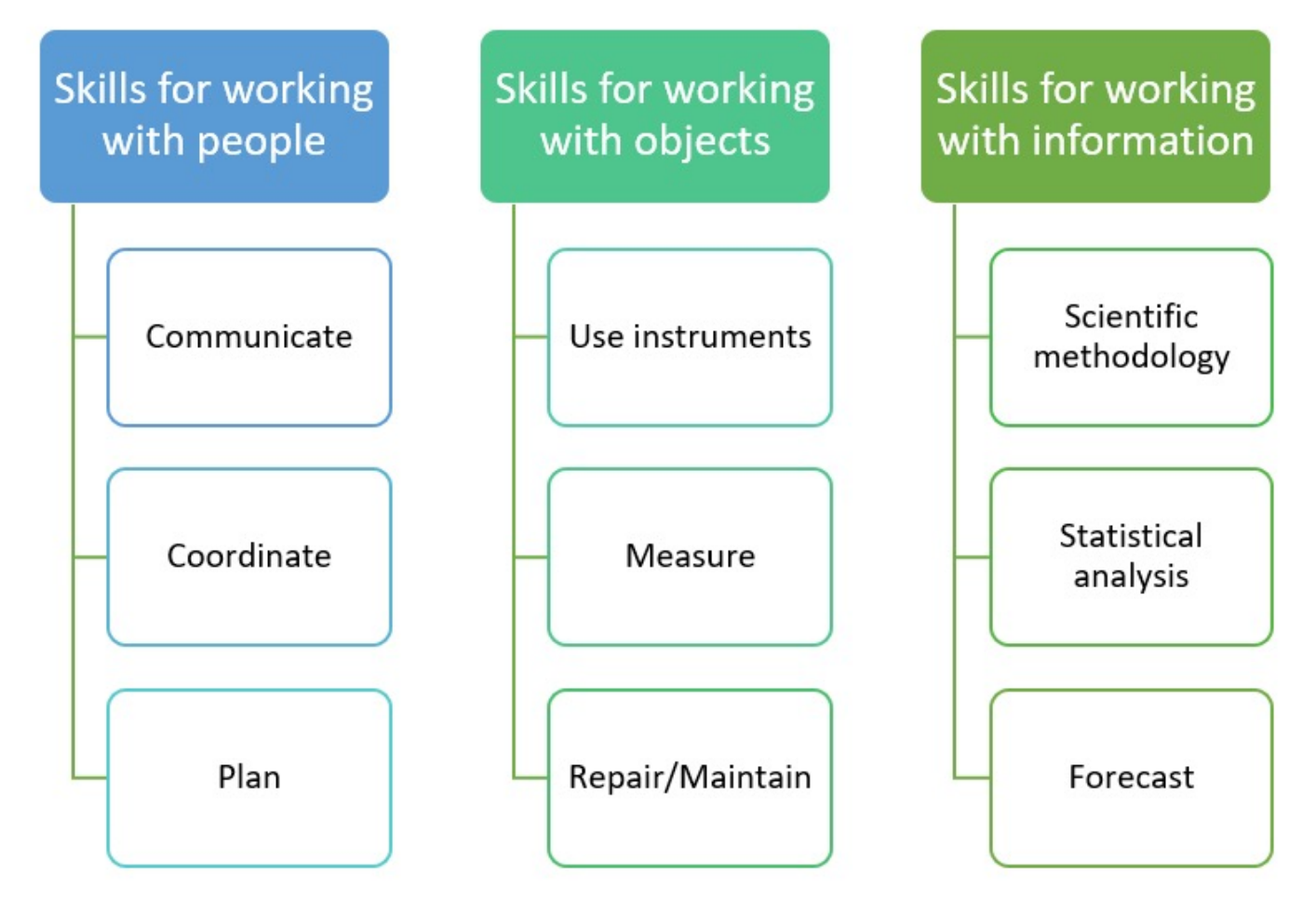
FIGURE 1-11 *Emerging Big Data ecosystem*

[EMC Education services]

#1.2 Scopul cursului

Antrenarea abilităților necesare în știința datelor

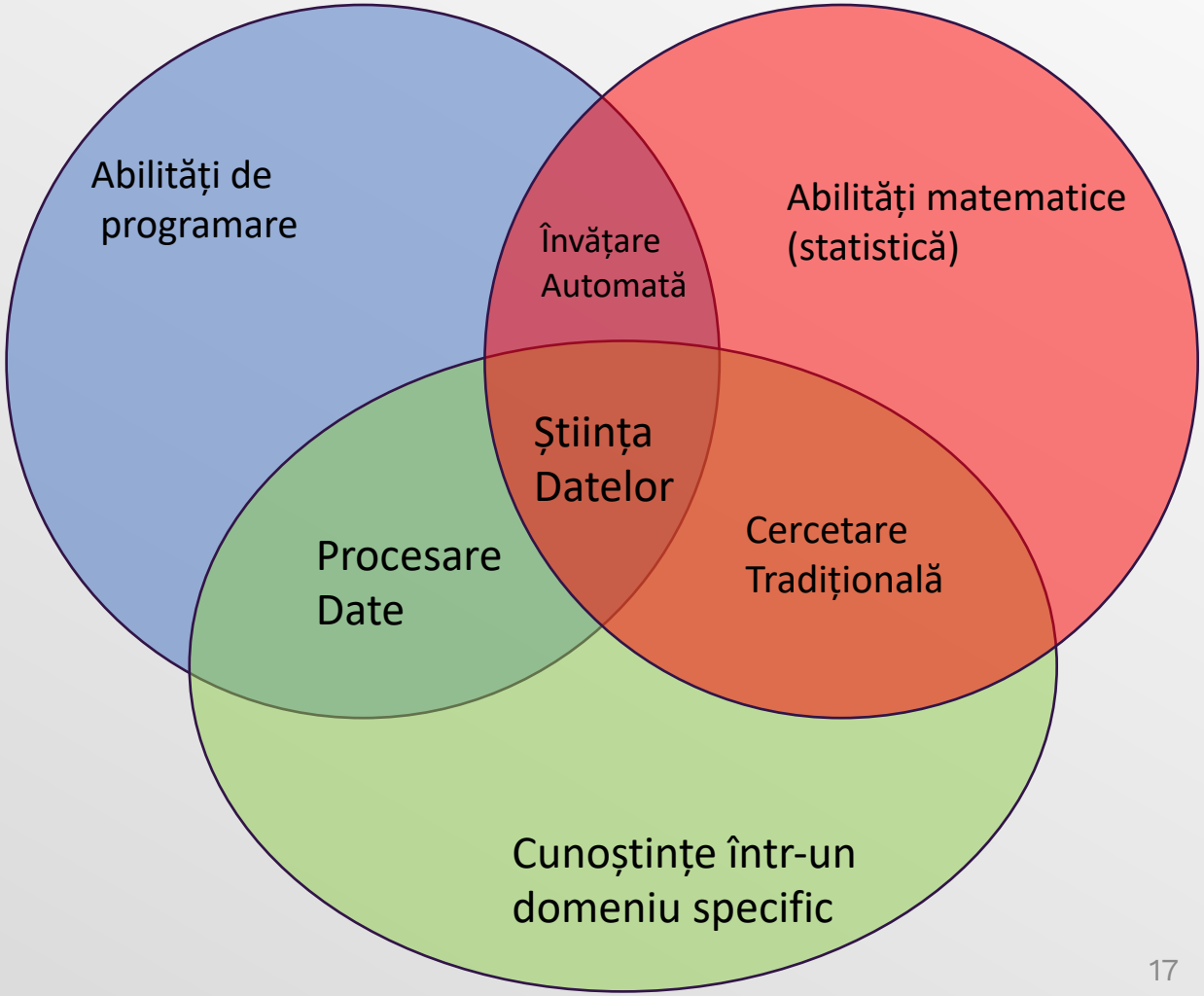
MIT Career Handbook



[MIT]

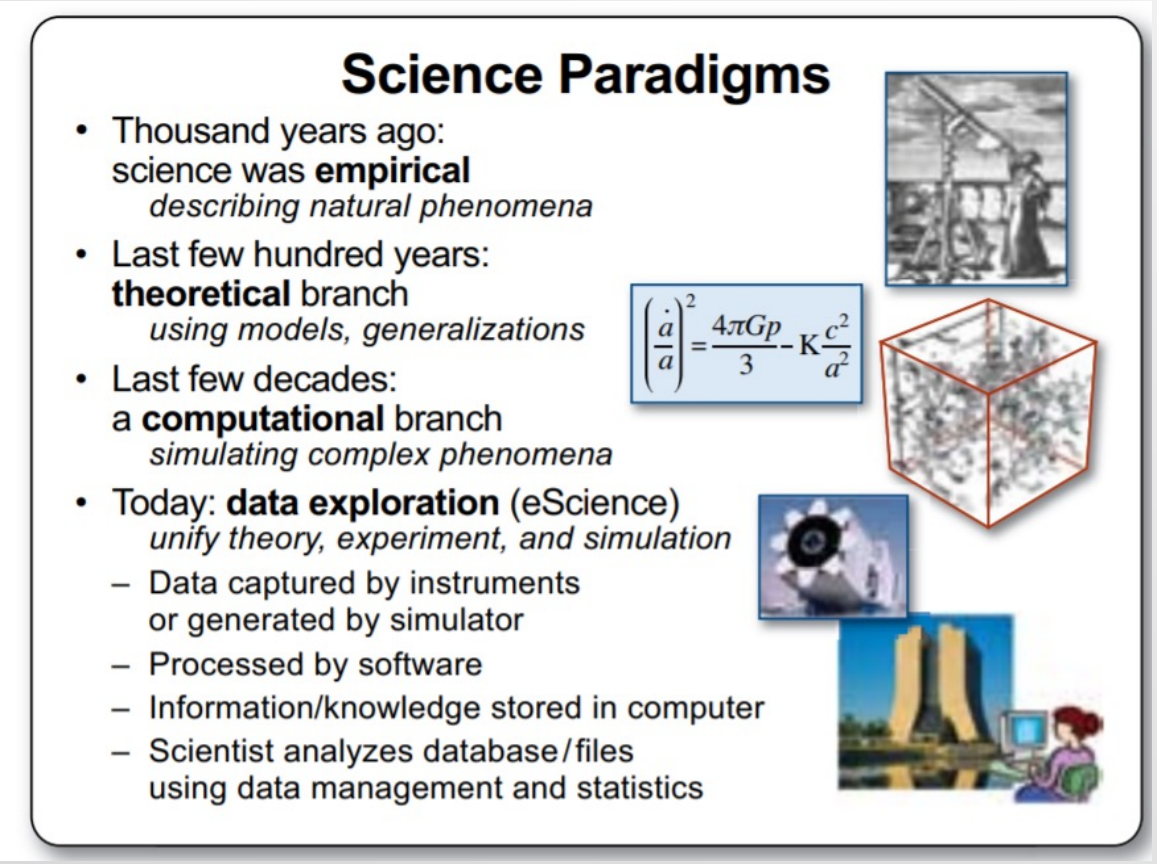
Abilități specifice SD

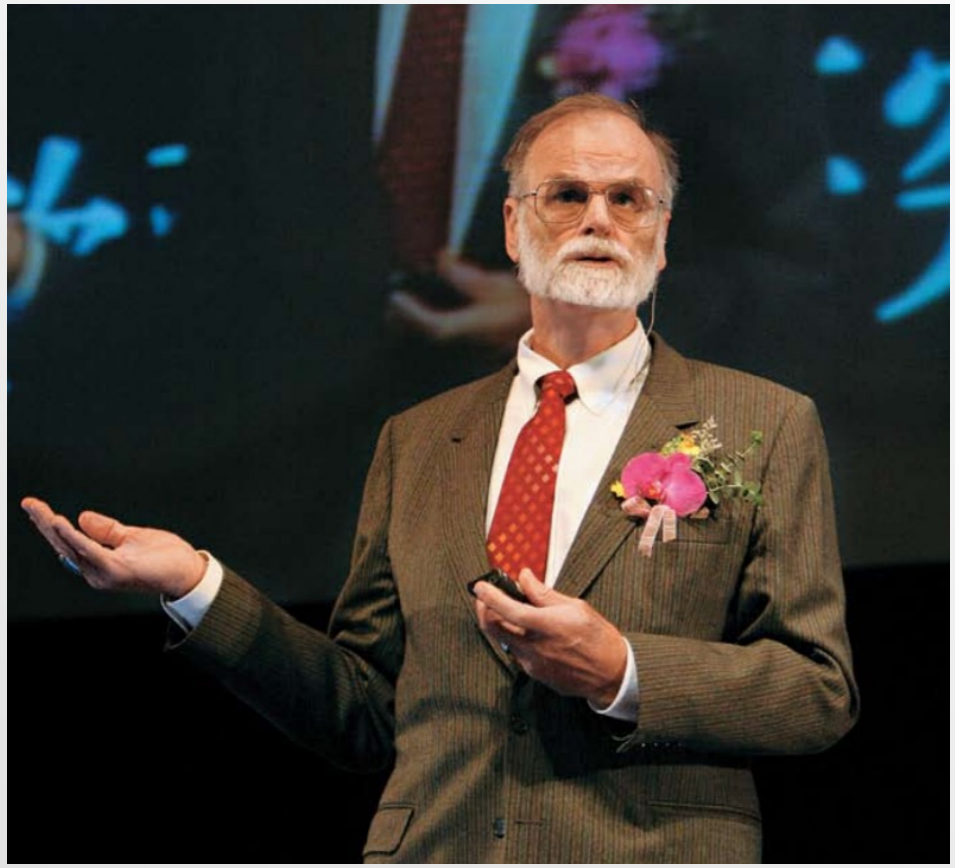
- Design-ul cercetării
  - Formulare a întrebărilor
  - Colectarea & analiza a datelor
  - Definirea modelului
- Automatizarea analizei
  - Programare
  - Procesarea datelor
- Comunicare
  - Formularea diagnozei, deciziei, predicției etc
  - Comunicarea cu beneficiarii



Importanța Științei Datelor – personalități

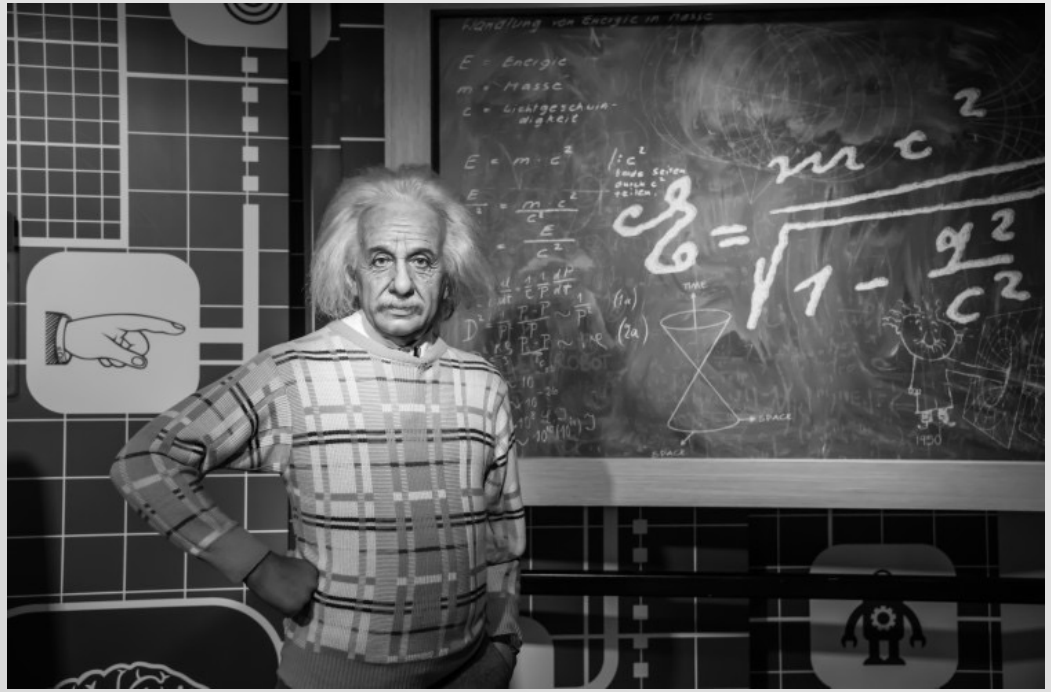
- Daryl Morey [Wiki, Interview]
  - MIT Sloan Sports Analytics.
  - Directorul general al Houston Racket (baschet)
- Mircea Lucescu
  - Antrenor român de fotbal [Wiki]
- Nate Silver [Wiki, FiveThirtyEight]
  - Predicții cu ajutorul datelor publice
  - Cine o să câștige alegerile ?

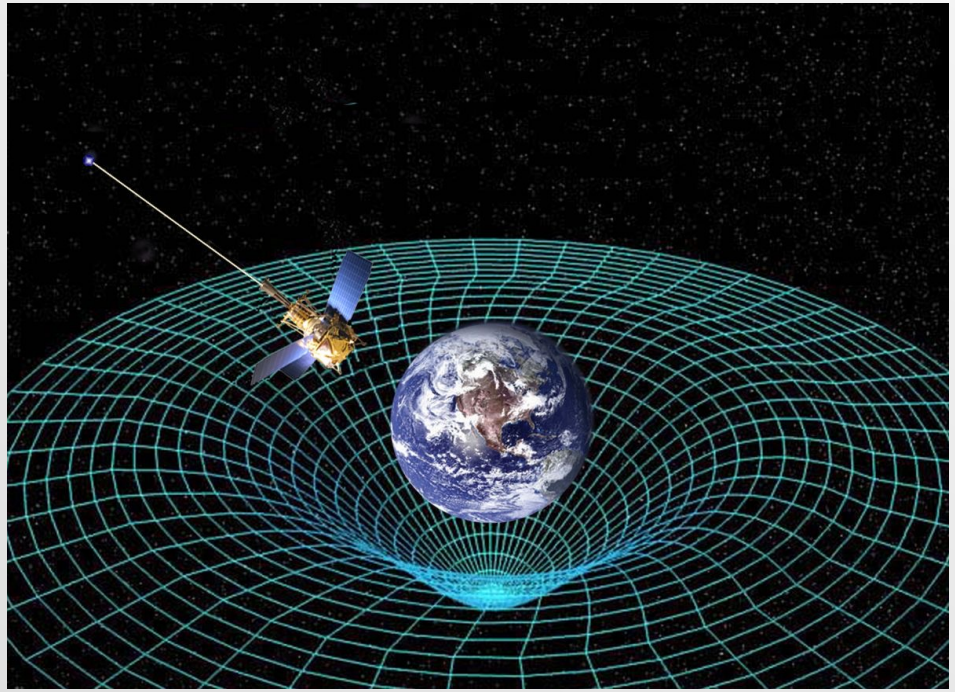




Jim Gray's Fourth paradigm of science [Hey et al.]

O nouă Teorie a Relativității?





Vezi: [Einstein's Relativity and Everyday Life, de Clifford M. Will]

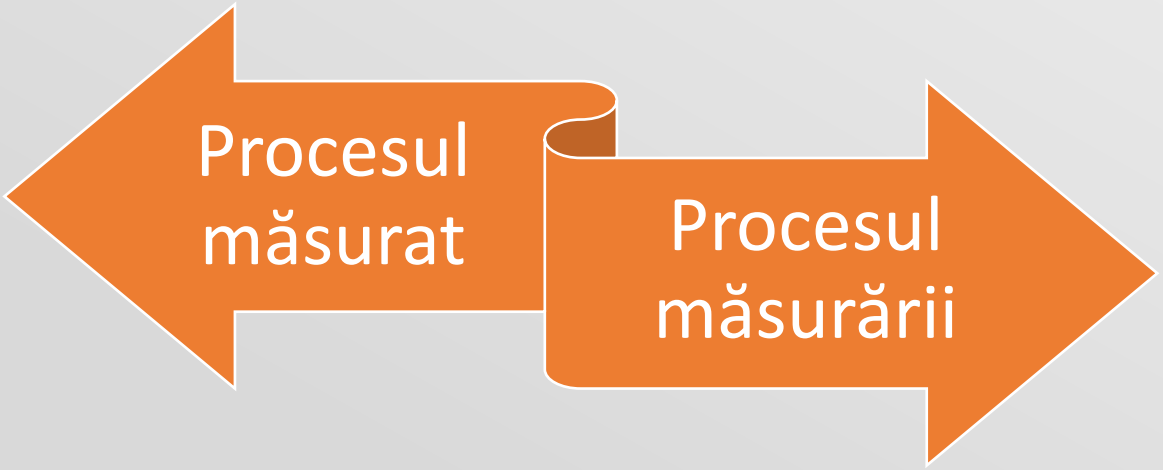
#1.3 Măsurarea și colectarea datelor

Măsurare vs. colectarea datelor

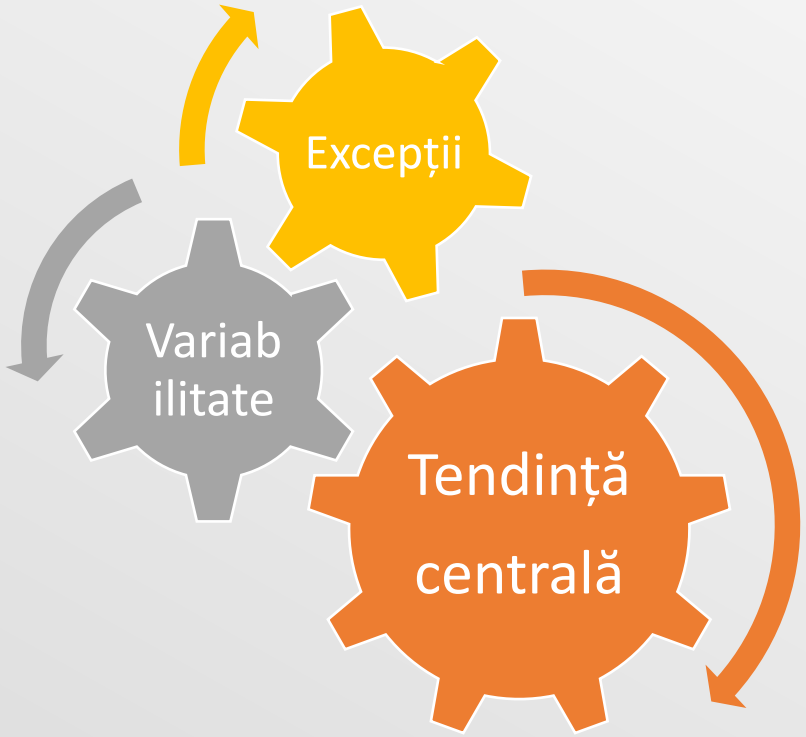
- Măsurarea: transformarea **fenomenelor** în **variabile**
  - Înălțimea → un număr de cm
  - Greutatea → un număr de kg
  - Temperatura → un număr de grade
- Colectarea datelor: organizarea & stocarea datelor din:
  - Măsurători proprii
  - Măsurători din surse diferite
- Heterogenitatea datelor: fiecare sursă poate avea propria metodologie

Două competențe centrale în DS

Să facem diferența între:



Să ținem cont mereu de:



Date: Observarea distribuției valorilor

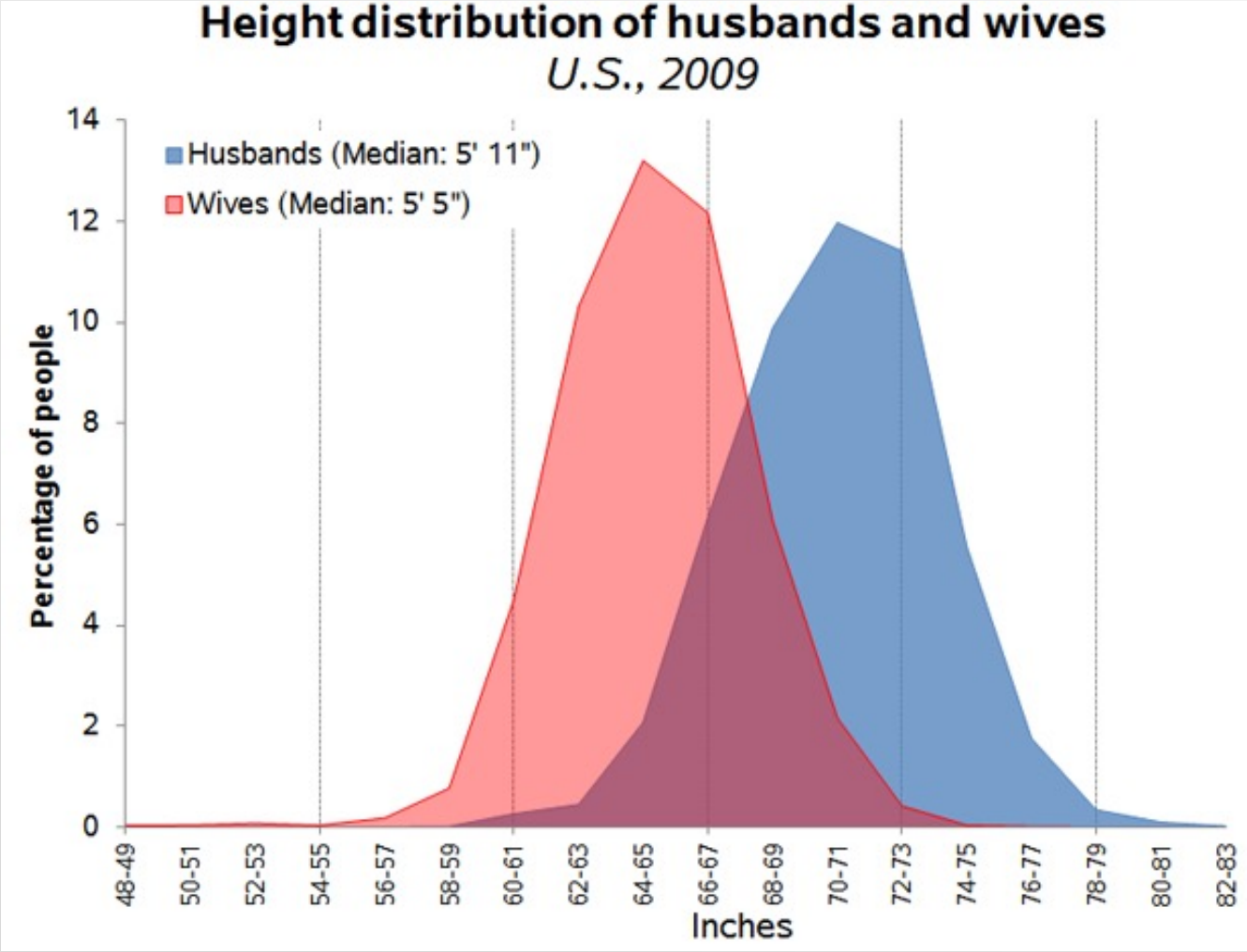
Stereotipuri vs. măsurare

- Stereotipurile sunt adesea și corecte, și greșite
  - Bărbații sunt mai înalți decât femeile
  - Femeile sunt mai puțin agresive decât bărbații
- Prin măsurare putem vedea:
  - Tendința centrală a fenomenului (cum este în linii mari)
  - Variabilitate
    - Corelații cu alte variabile: cauze?
    - Variabilitatea variabilității: evoluții, hărți
  - Excepții, outliers

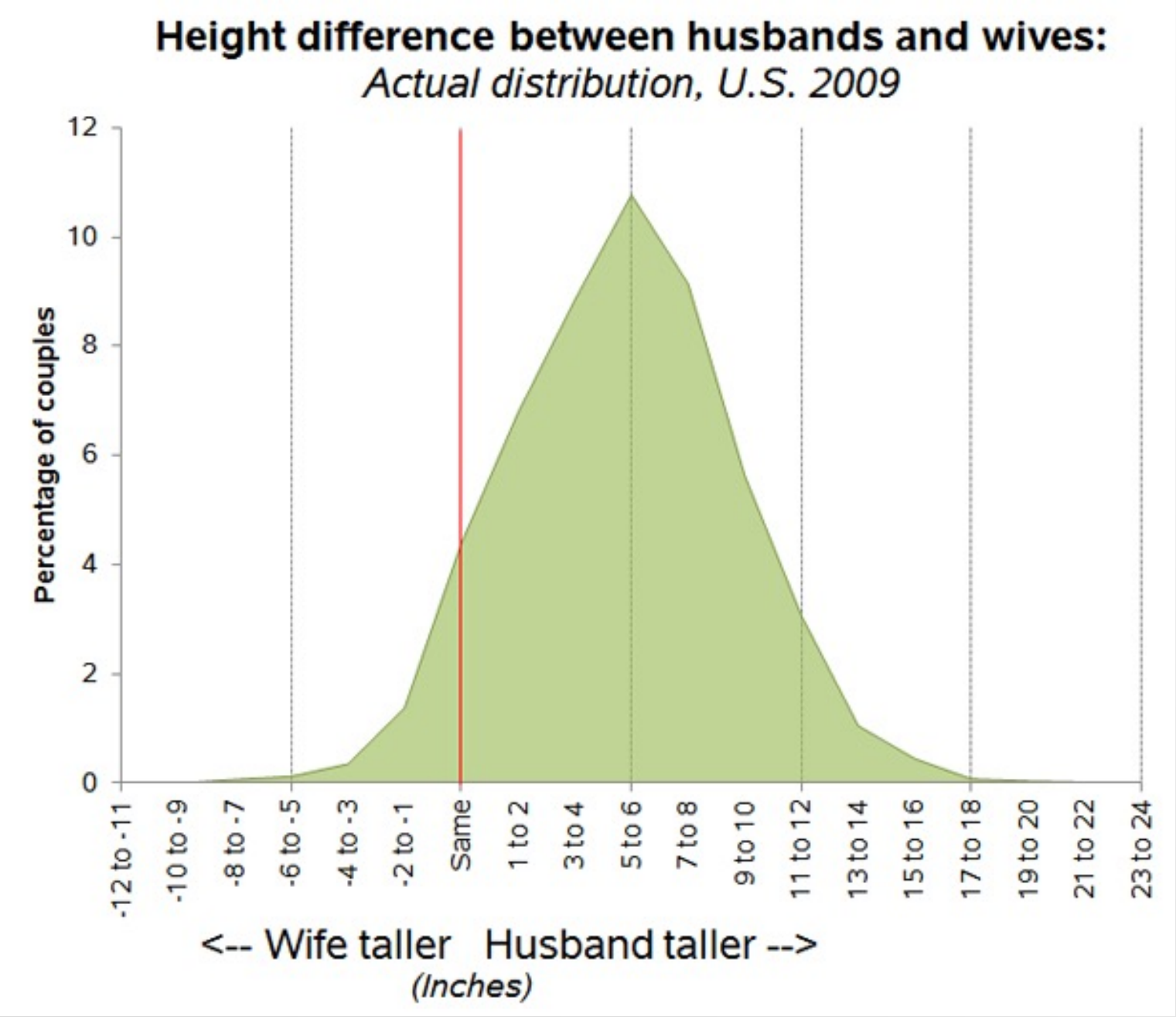
Procese reale vs. măsurare

- Întotdeauna trebuie să înțelegem:
  - Procesul măsurat
  - Procesul măsurării
- Putem identifica:
  - Posibile erori de gândire, stereotipuri
  - Relații de corelație și cauzalitate
  - Tipologii
  - Posibile erori de măsurare

Variabile și distribuții

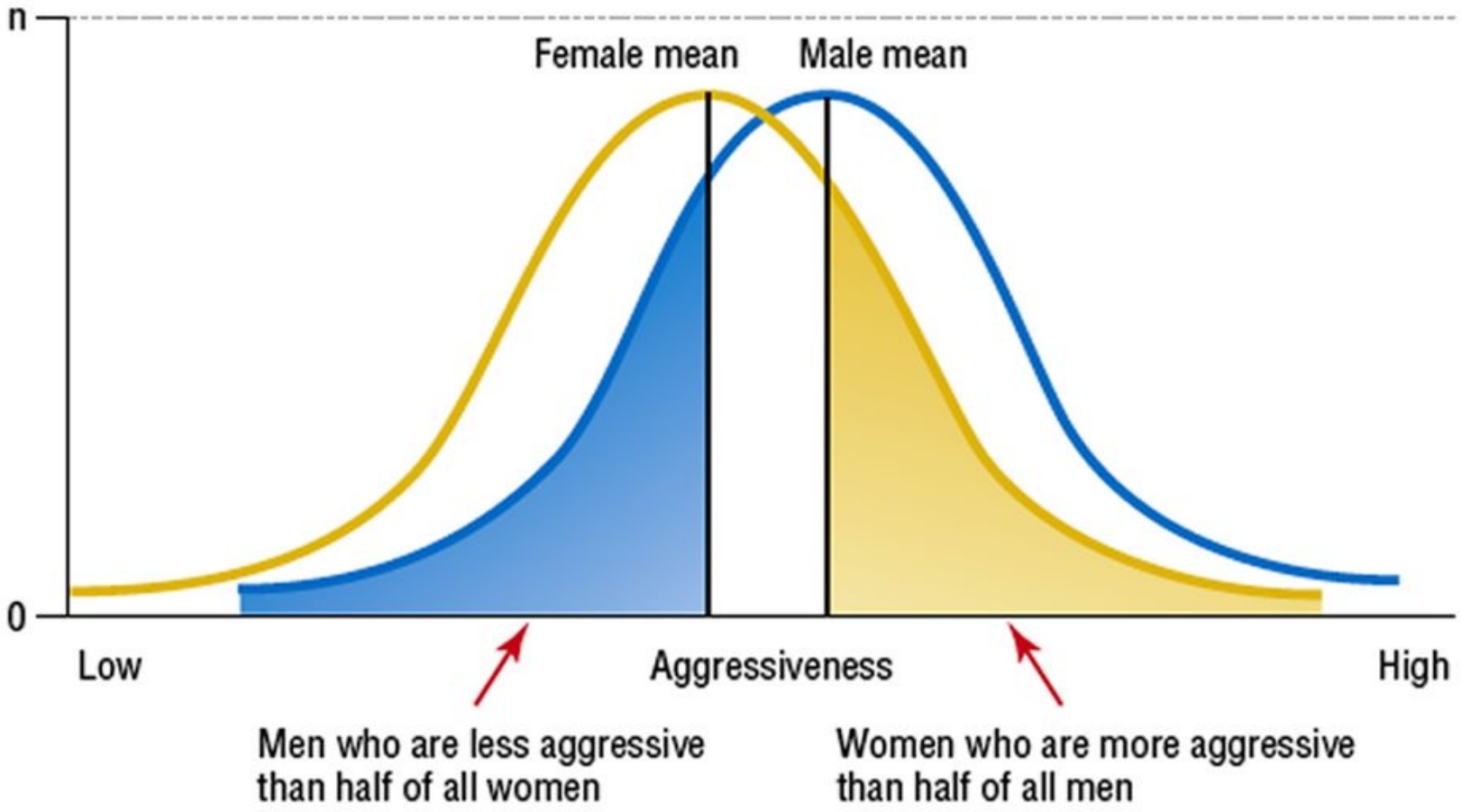


The Atlantic



The Atlantic

*Distribution of Aggressiveness among Men and Women

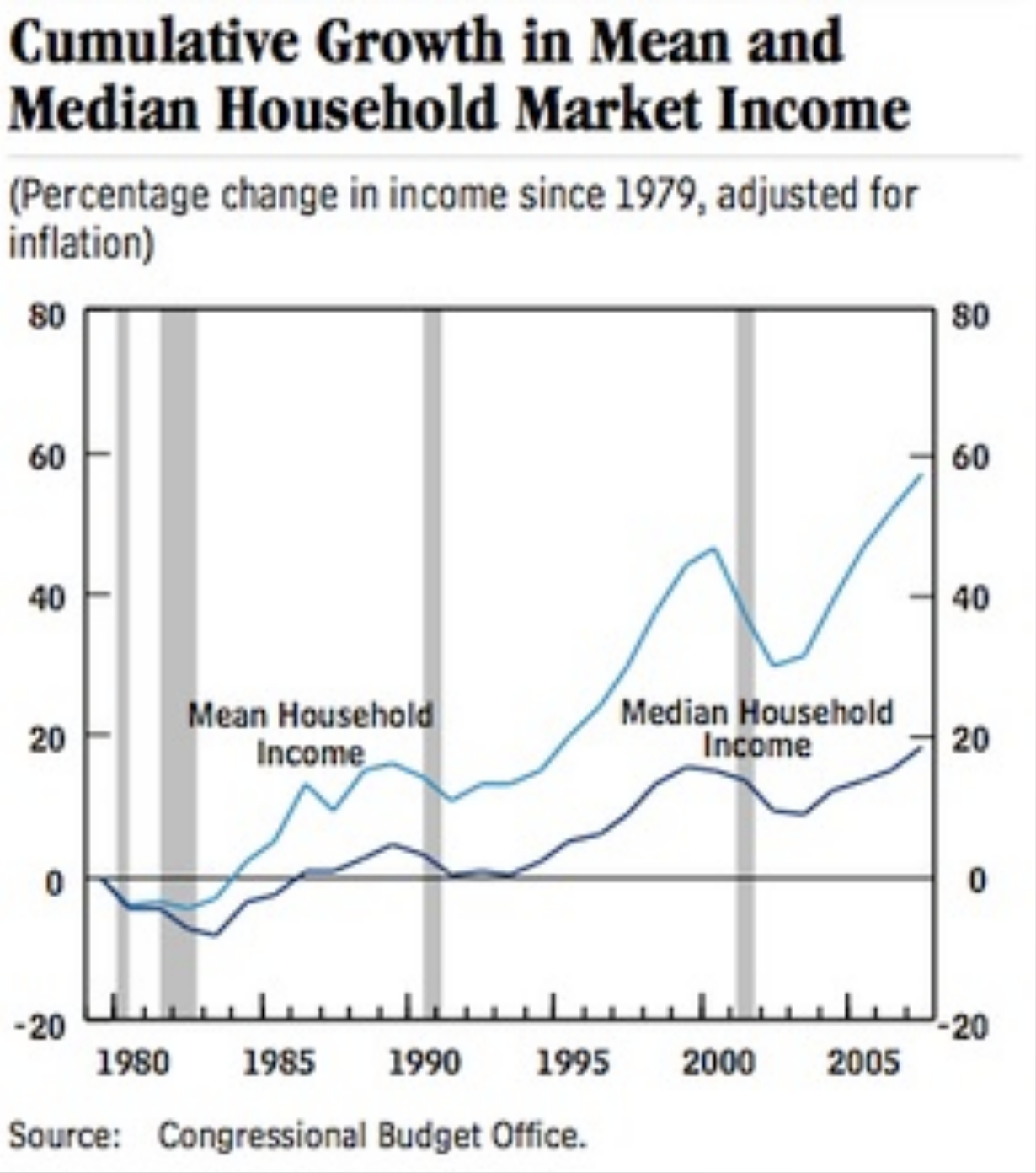


Sursă: Miller

Tendință centrală și variabilitate

- Fie $X = \{x_1, x_2, \dots, x_n\}$ *și* $n = |X|$
- Tendința centrală
  - **Medie**: $\mu_X = \frac{\sum_{i=1}^{n} x_i}{n}$
  - **Mediană**: separă 50% cazuri la stânga, 50% cazuri la dreapta
  - **Mod**: este cea mai frecventă valoare
  - **Percentile (cuartile, cvintile etc)**: separă x% din cazuri
- Variabilitatea: cât de mult variază față de medie
  - Variația $\sigma_X^2 = \frac{\sum_{i=1}^{n}(x_i - \mu_X)^2}{n}$
  - Deviația standard $\sigma_X = \sqrt[2]{\sigma_X^2}$

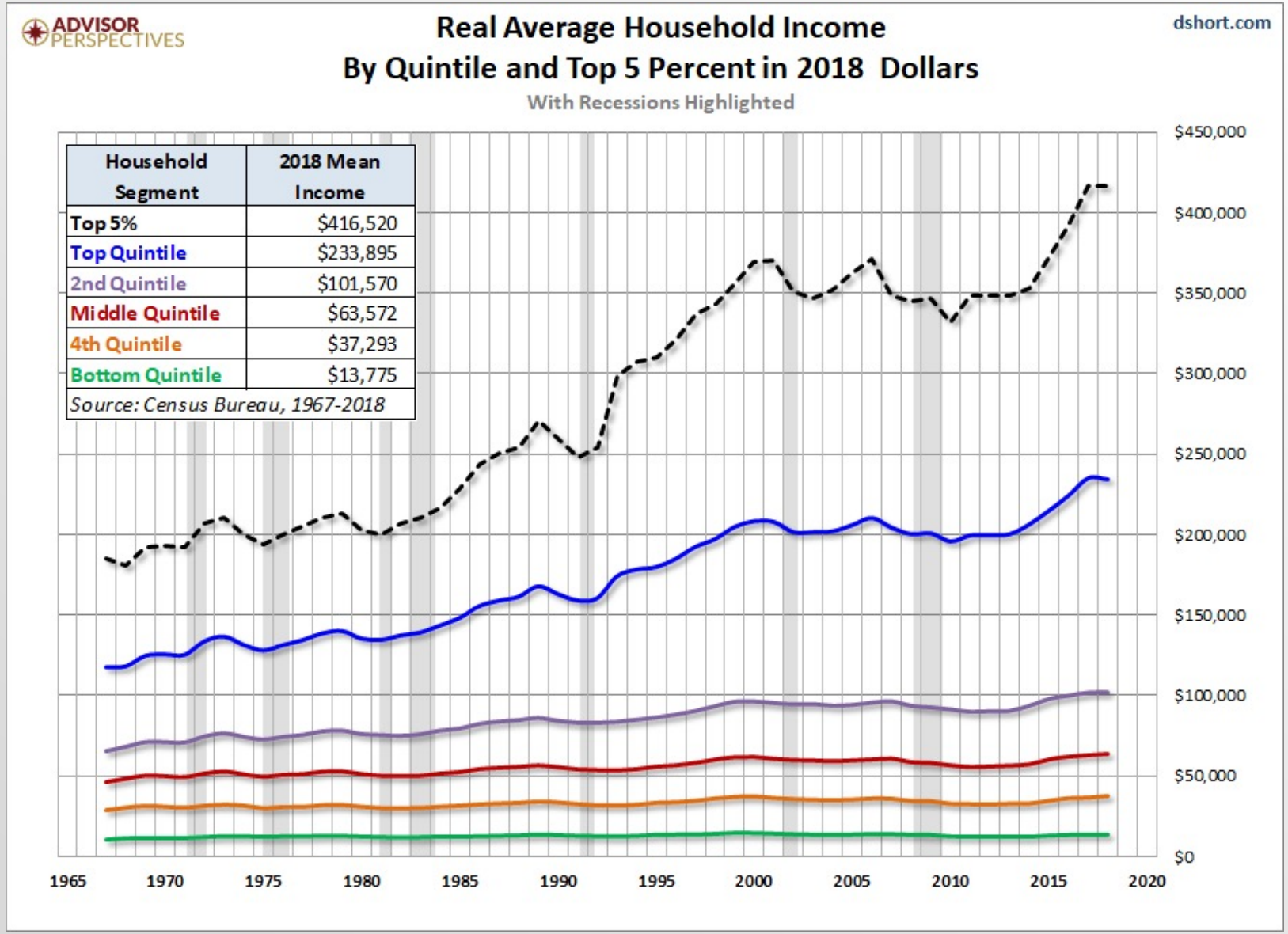
Tendința centrală: Medie vs. mediană



The Atlantic

Percentile:

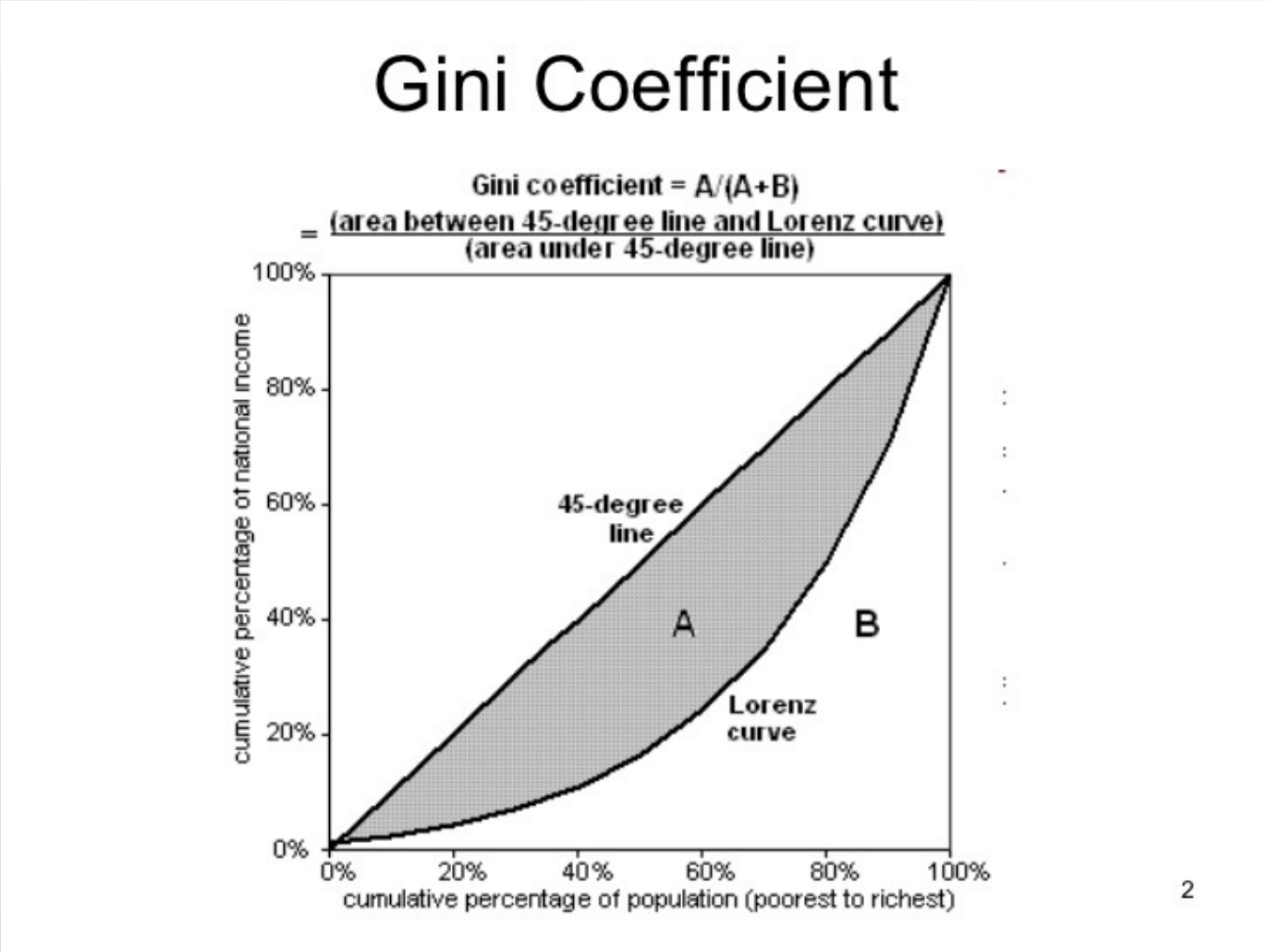
Înțelegerea dinamicilor reale a creșterii venitului



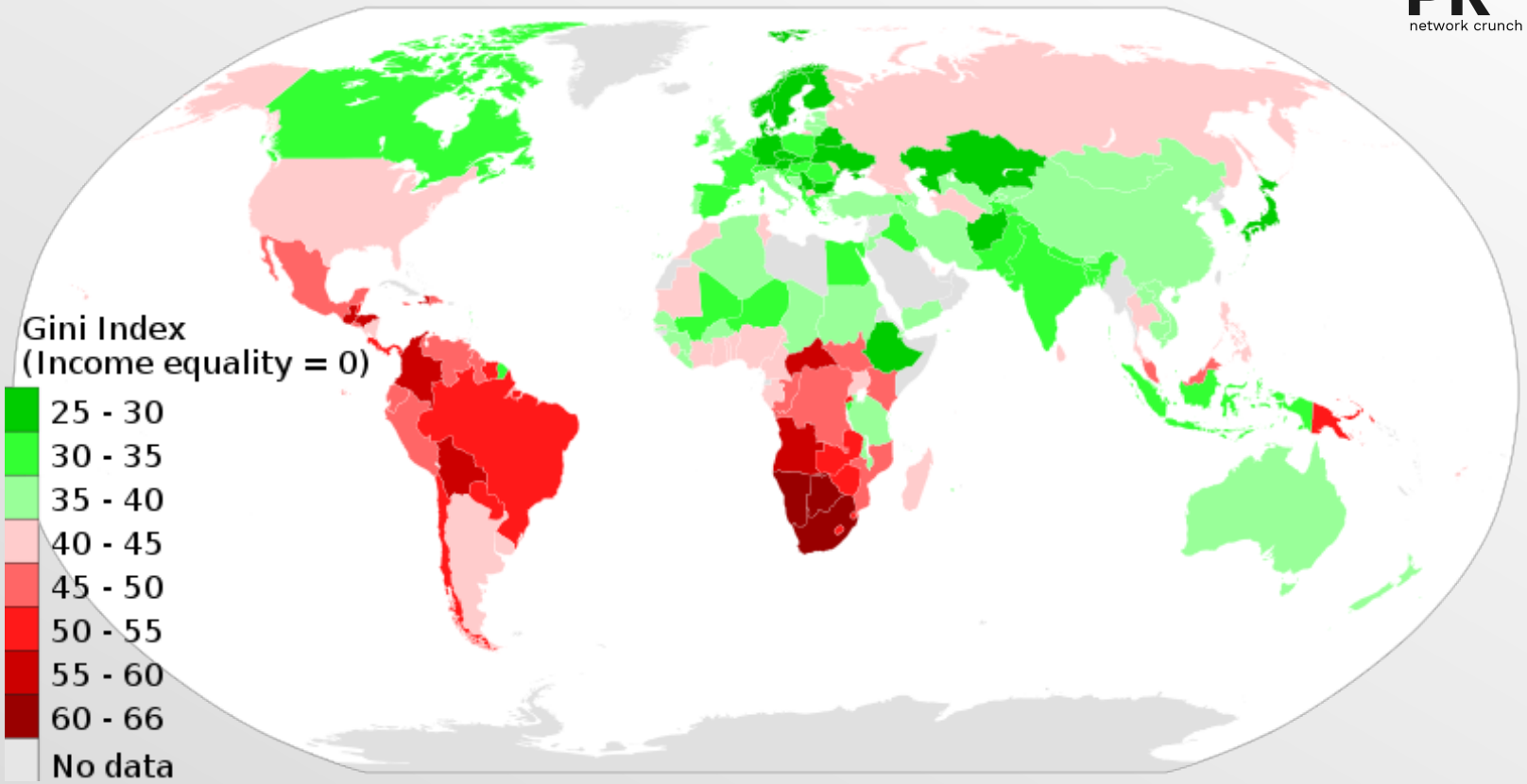
Sursă

Variabilitatea averii:

Coeficientul Gini



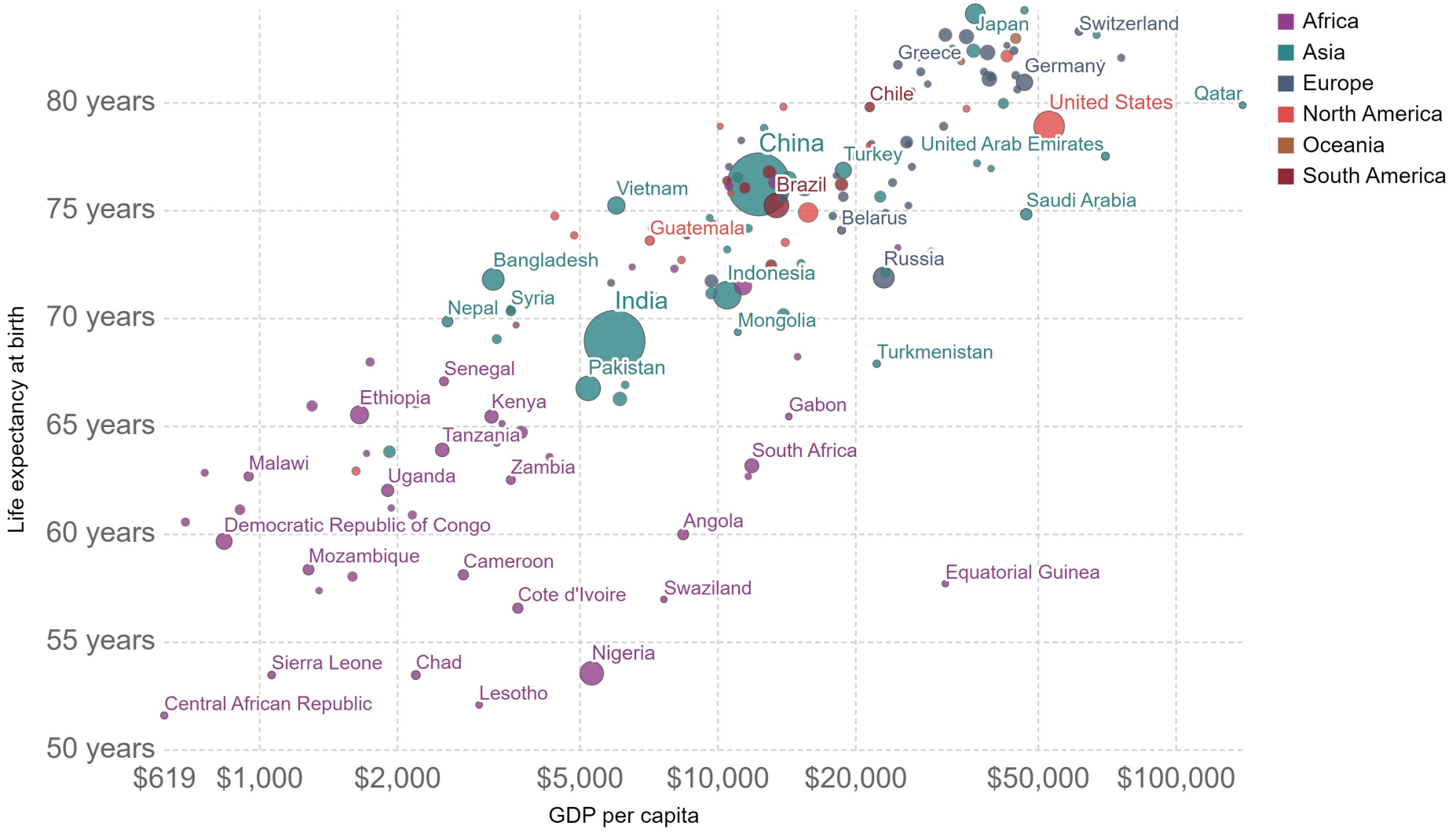
Variabilitatea variabilității averii!



Wikimedia

Life expectancy vs. GDP per capita, 2016

GDP per capita is measured in 2011 international dollars, which corrects for inflation and cross-country price differences.



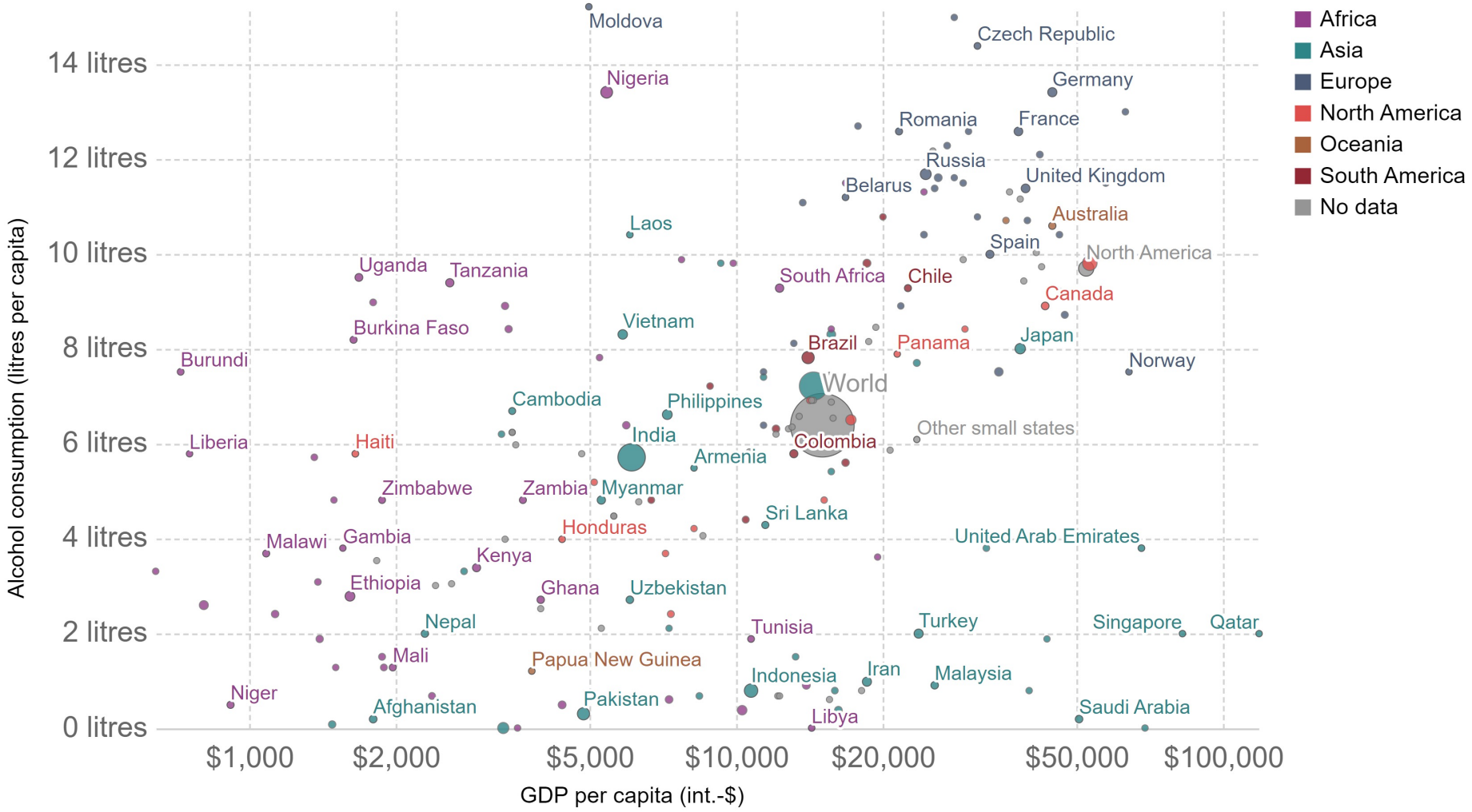
Source: Clio-Infra & UN Population Division ; Maddison Project Database (2018)

OurWorldInData.org/life-expectancy • CC BY

Sursă

Alcohol consumption vs. GDP per capita, 2016

Average annual alcohol consumption measured in liters of pure alcohol per person (15 years of age or older). Gross domestic product (GDP) per capita measured is adjusted for differences in price levels between countries (measured in international-$).



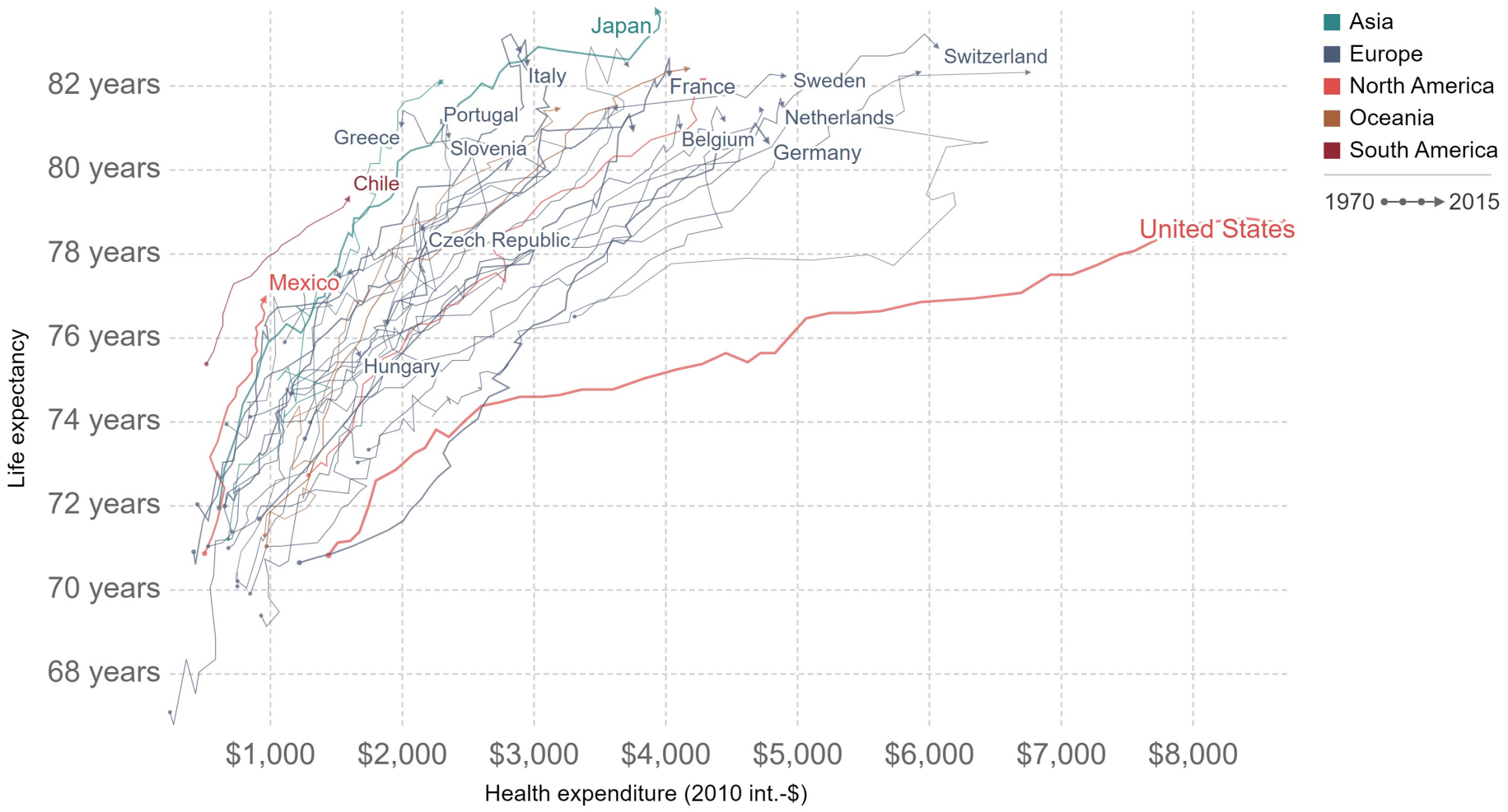
Source: World Bank, Population (Gapminder, HYDE(2016) & UN (2019)), Our World In Data

CC BY

Sursă

Life expectancy vs. health expenditure, 1970 to 2015

Health financing is reported as the annual per capita health expenditure and is adjusted for inflation and price level differences between countries (measured in 2010 international dollars).



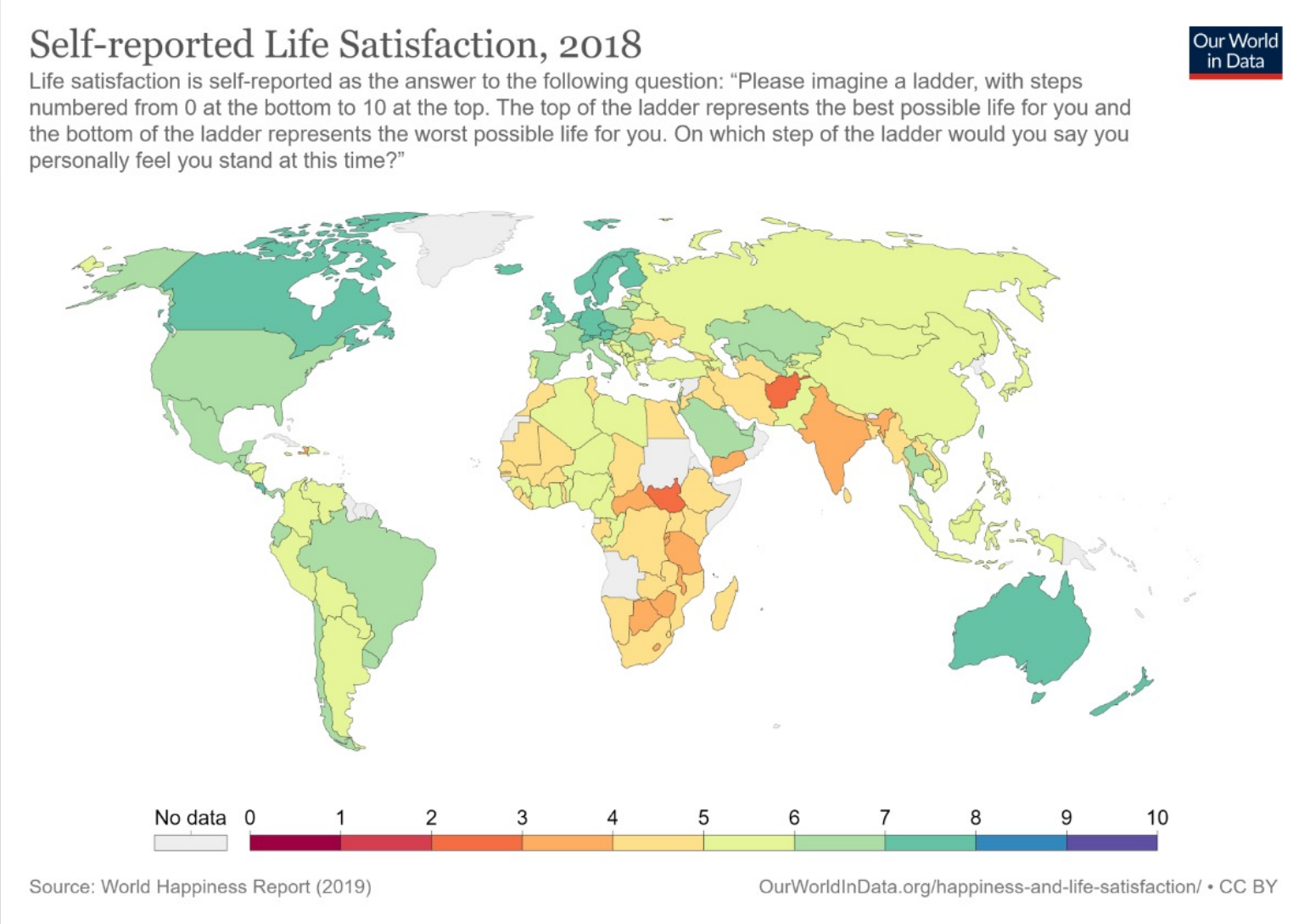
Source: World Bank, Health Expenditure and Financing - OECDstat (2017), Population (Gapminder, HYDE(2016) & UN (2019)), Our World In Data
OurWorldInData.org/the-link-between-life-expectancy-and-health-spending-us-focus • CC BY

Sursă

Niveluri de măsurare

- Nivelul **numeric**
  - Valorile sunt numere
  - Permite calcul algebric (medie)
- Nivelul **ordinal**: satisfacția cu calitatea aerului, percepția corupției
  - Valorile sunt cuvinte: foarte mult, mult, puțin, foarte puțin (scala Likert)
  - Valorile pot fi ordonate (permite calculul medianei și al modului)
  - Uneori traduse în numere
- Nivelul **nominal**: județul, numele de familie, zodia
  - Valorile sunt cuvinte
  - Valorile nu pot fi ordonate
  - Permite numărări, clasificări, dar nu operații algebrice (modul)

Măsurare ordinală: satisfacția cu viața



Sursă

#1.4 Erori de măsurare

Ce fel de erori de măsurare?

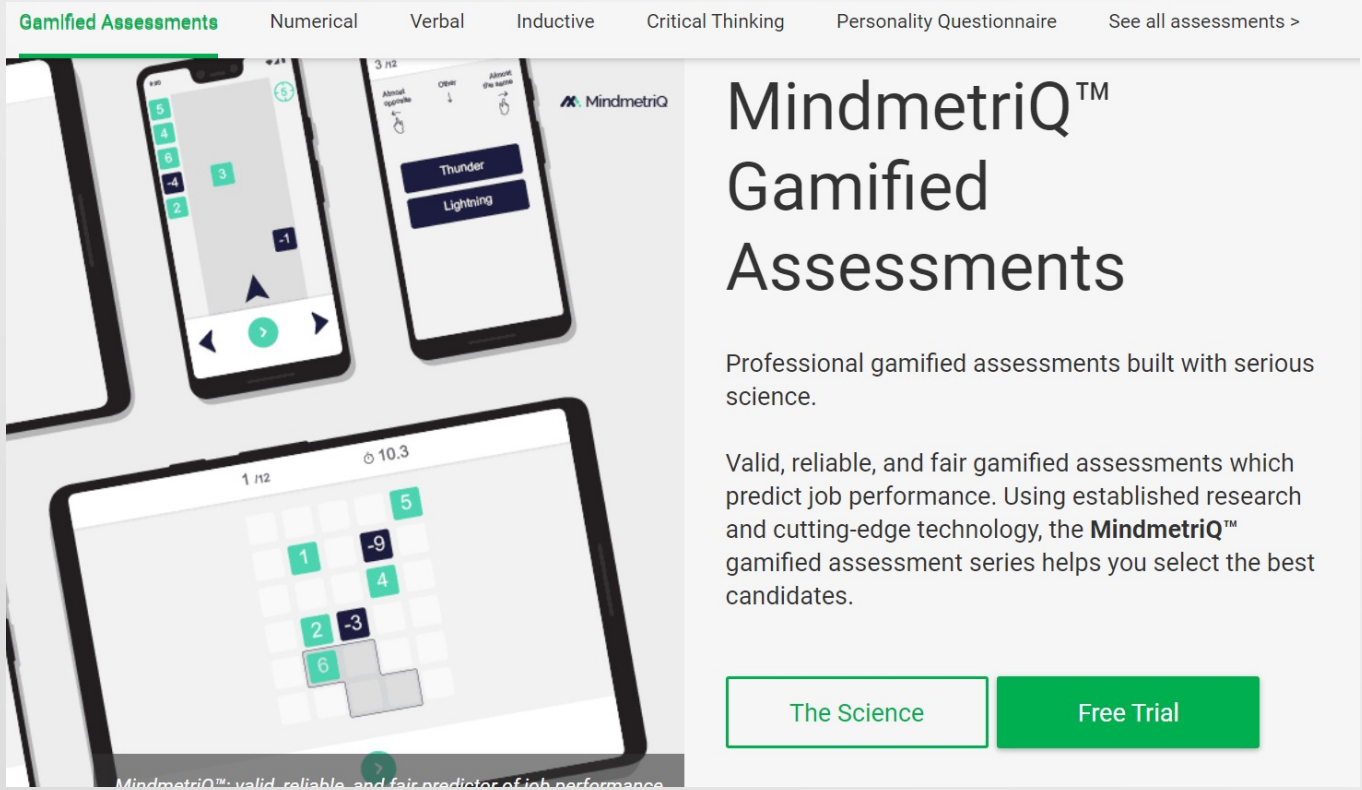
- Venitul măsurat prin declarațiile depuse la Fisc
- Temperatura măsurată punând mâna pe frunte
- Popularitatea unui candidat politic măsurată prin sondaje de opinie
- Talentul unui copil măsurat întrebând părinții
- Riscul de accident măsurat prin teama pasagerilor într-un avion

Erori: precizia măsurării

- En. *accuracy*
- Valoarea măsurată - valoarea reală =?
- Tipuri de erori în funcție de direcția lor:
  - Erori **aleatoare**
  - Erori **sistematice**: dezirabilitatea socială
    - Orientarea sexuală
    - Preferințele politice
- Măsurarea timpului:
  - Ochiometric, clepsidră: erori aleatoare
  - Un ceas care rămâne în urmă: erori sistematice

Erori: Validitatea măsurării

- En. *validity*
- **Ce vrei** să măsori vs. **ce măsori de fapt**?
- Ex.:
  - Măsoară prețul calitatea produsului?
  - Măsoară un test gamificat abilitățile candidatului?



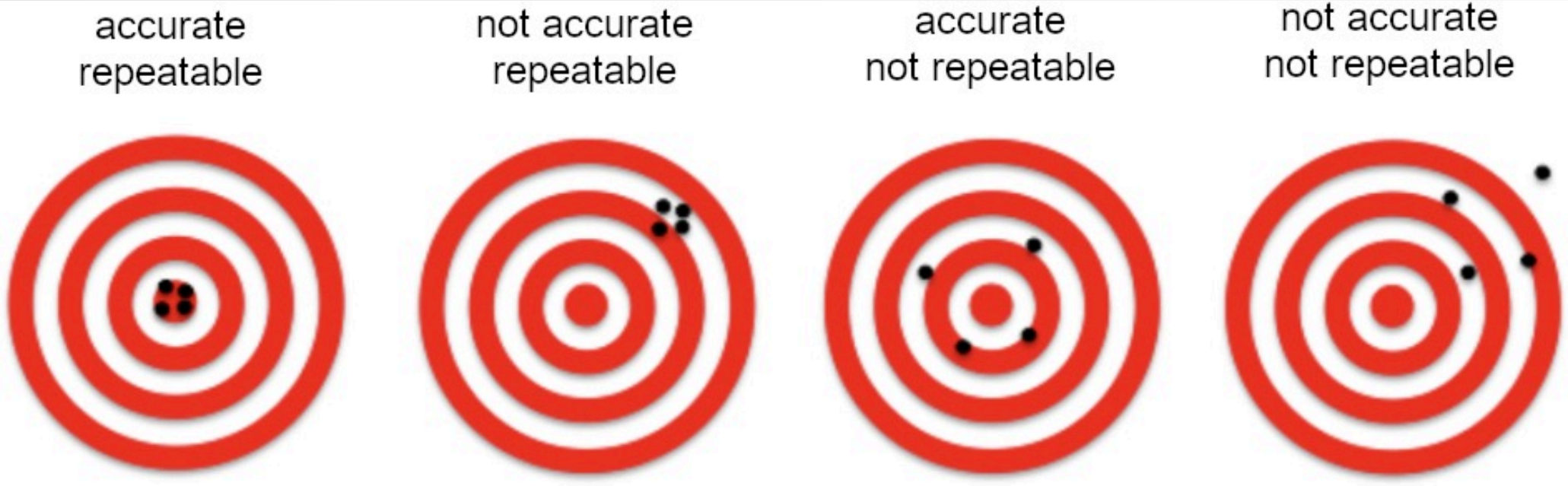
Erori: Fidelitatea măsurării

- En. *reliability*
- Cât de **repetabil** este rezultatul?
- Măsurători repetate pentru același nivel dau aceeași valoare?
  - Măsurarea temperaturii scoțând capul pe geam
  - Măsurarea temperaturii cu un termometru

Măsurarea zodiei

- Foarte precis
- Foarte repetabil
- Cât de valid?

Scenarii



Temperatura [termometrul]
- Erori aleatorii reduse

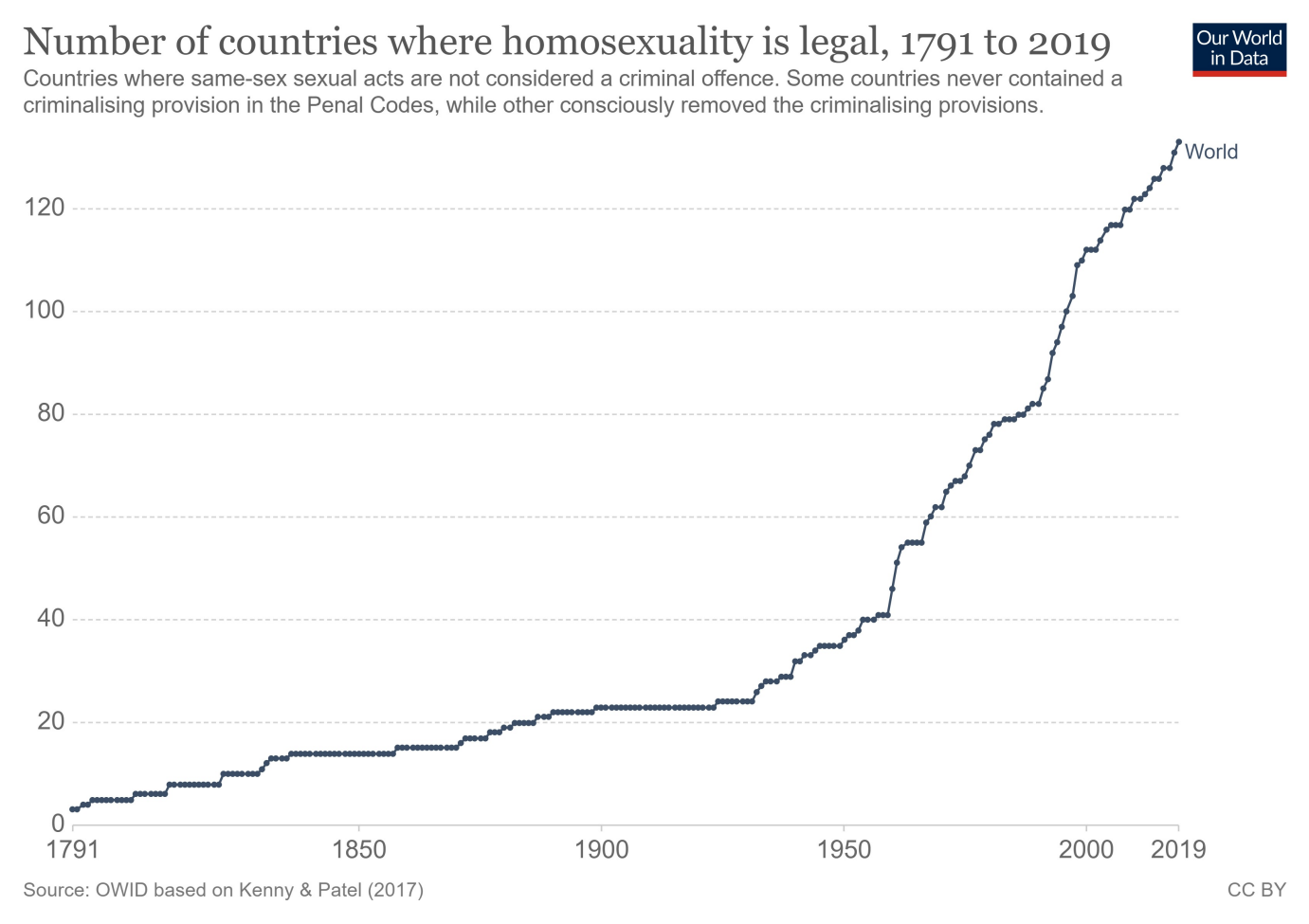
Ora [ceas care o ia înainte]
Orientarea sexuală [chestionar]
- Erori sistematice dar estimări stabile

Ora
[capul pe geam]
- Erori aleatorii mari

Propria inteligență [auto-estimare]
- Erori sistematice și valori instabile

Sursă

Măsurarea orientării sexuale prin chestionar



Facebook vs. variabilitate culturală

- Trăsături personale pot fi inferate pe baza like-urilor Facebook [Kosinski et al.]
  - Validare pe un lot de voluntari din SUA
- Cât de valide sunt inferențele pentru utilizatori din România?

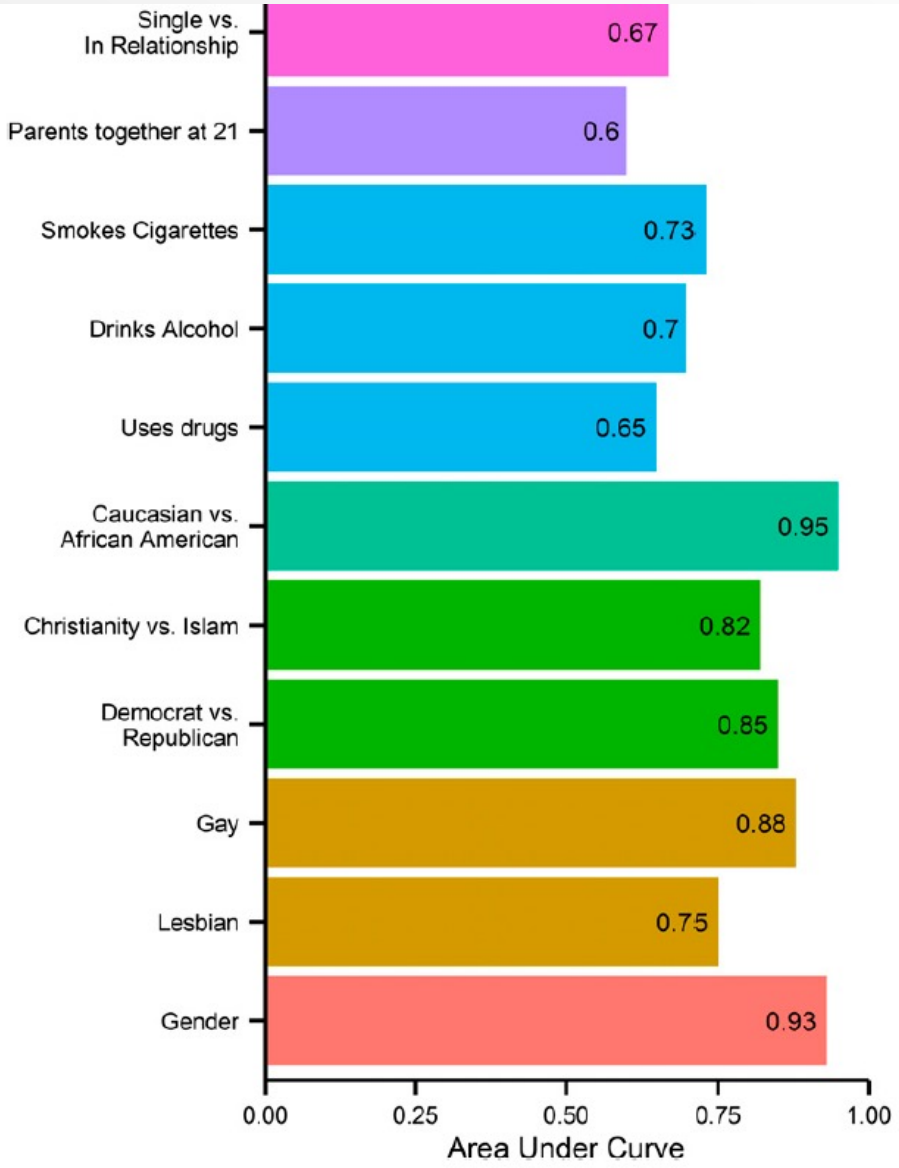


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

Pe scurt despre știința datelor

- Abilități esențiale privind lucrul cu informații
- Abilități tot mai cerute datorită avansului AI și Big Data
- Două competențe cheie: să ții cont simultan de...
  - Procesul măsurat și procesul măsurării
  - Tendința centrală, variabilitate și excepții
- Tendința centrală e estimată prin: medie, mediană, mod
- Variabilitatea: variația, deviația standard
- Erorile de măsurare pot privi:
  - Precizia: sistematice / aleatoare
  - Validitatea
  - Fidelitatea (repetabilitatea)

Bibliografie

- [1] - Peng, Roger D., and Elizabeth Matsui. "The art of data science." A Guide for Anyone Who Works with Data. Skybrude Consulting, LLC (2015).
- [2] - Coale, Ansley J. "The history of the human population." Scientific American 231.3 (1974): 40-51.
- [3] - https://www.statista.com/statistics/871513/worldwide-data-created/
- [4] - Guo, Huadong. "Big Earth data: A new frontier in Earth and information sciences." Big Earth Data 1.1-2 (2017): 4-20.

Bibliografie

- [5] http://www.johnstonsarchive.net/other/worldpop.html
- [6] - Gustafson, John L. *The End of Error: Unum Computing*. CRC Press, 2017.
- [7] - Gustafson, John L., and Isaac T. Yonemoto. "Beating floating point at its own game: Posit arithmetic." Supercomputing Frontiers and Innovations 4.2 (2017): 71-86.
- [8] - Morris, Robert. "Tapered floating point: A new floating-point representation." *IEEE Transactions on Computers* 100.12 (1971): 1578-1579.

Bibliografie

- [9] - Lakhina, Anukool, et al. "Structural analysis of network traffic flows." Proceedings of the joint international conference on Measurement and modeling of computer systems. 2004.
- [10] - Gu, Guofei, et al. "Botminer: Clustering analysis of network traffic for protocol-and structure-independent botnet detection." (2008): 139.
- [11] - Dekeris, Brunonas, Tomas Adomkus, and Aurelijus Budnikas. "Analysis of QoS assurance using weighted fair queueing (WQF) scheduling discipline with low latency queue (LLQ)." 28th International Conference on Information Technology Interfaces, 2006.. IEEE, 2006.
- [12] - Kuang, Liwei, et al. "A tensor-based big data model for QoS improvement in software defined networks." *IEEE Network* 30.1 (2016): 30-35.

Bibliografie

- [13] - Erman, Jeffrey, Martin Arlitt, and Anirban Mahanti. "Traffic classification using clustering algorithms." Proceedings of the 2006 SIGCOMM workshop on Mining network data. 2006.
- [14] - Le, Luong-Vy, et al. "Applying big data, machine learning, and SDN/NFV to 5G traffic clustering, forecasting, and management." 2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft). IEEE, 2018.
- [15] - Cisco. "Cisco visual networking index: Global mobile data traffic forecast update 2017–2022." Update 2017 (2019): 2022.
- [16] - https://insights.stackoverflow.com/survey/2020

Bibliografie

- [17] Hey, Tony, Stewart Tansley, and Kristin M. Tolle. "Jim Gray on eScience: a transformed scientific method." (2009).
- [18] Chiang, C. 2018. Defining the terms: structured vs unstructured data. https://www.igneous.io/blog/structured-data-vs-unstructured-data
- [19] EMC Education Services. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Wiley, 2015.
- [20] Watson, J., Whiting, P. F., & Brush, J. E. (2020). Interpreting a COVID-19 test result. BMJ, 369.
- [21] Kaufman, S. B., Yaden, D. B., Hyde, E., & Tsukayama, E. (2019). The light vs. dark triad of personality: Contrasting two very different profiles of human nature. Frontiers in psychology, 10, 467.
- [22] Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the national academy of sciences, 110(15), 5802-5805.

SYSTEMS LABORATORY

Cursul DS02

Corelația. Erori de modelare și de eșantionare

PR network crunch

Structura cursului

- Introducere: De ce DS?
- Scopul cursului
- Măsurarea și colectarea datelor
- Erori de măsurare
  - Precizia
  - Validitatea
  - Fidelitatea
- Corelația
- Erori de modelare
  - Corelații iluzorii
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erori de eșantionare
- Tehnologiile datelor: programare vs. Învățare automată
- Procesul învățării automate
- Predicție și cunoaștere: analiza datelor
  - Măsurare și reducerea dimensionalității
  - Clasificare (cluster)
  - Cauzalitate și regresie
  - Predicții și serii de timp
- Predicția ca premisă pentru acțiune
- Predicția ca produs economic

DS01: Pe scurt despre știința datelor

- Abilități esențiale privind lucrul cu informații
- Abilități tot mai cerute datorită avansului AI și Big Data
- Două competențe cheie: să ții cont simultan de...
  - Procesul măsurat și procesul măsurării
  - Tendința centrală, variabilitate și excepții
- Tendința centrală e estimată prin: medie, mediană, mod
- Variabilitatea: variația, deviația standard
- Erorile de măsurare pot privi:
  - Precizia: sistematice / aleatoare
  - Validitatea
  - Fidelitatea (repetabilitatea)

Cuprins

2.1 Corelația

2.2 Erori de modelare sau corelații iluzorii

- Sensul cauzării
- A fi corect din rațiuni false
- Corelații coincidentale
- Paradoxul lui Simpson
- Eroarea ecologică

2.3 Erori de eșantionare

2.4 Studiu de caz – Nobel și ciocolată

#2.1 Corelația

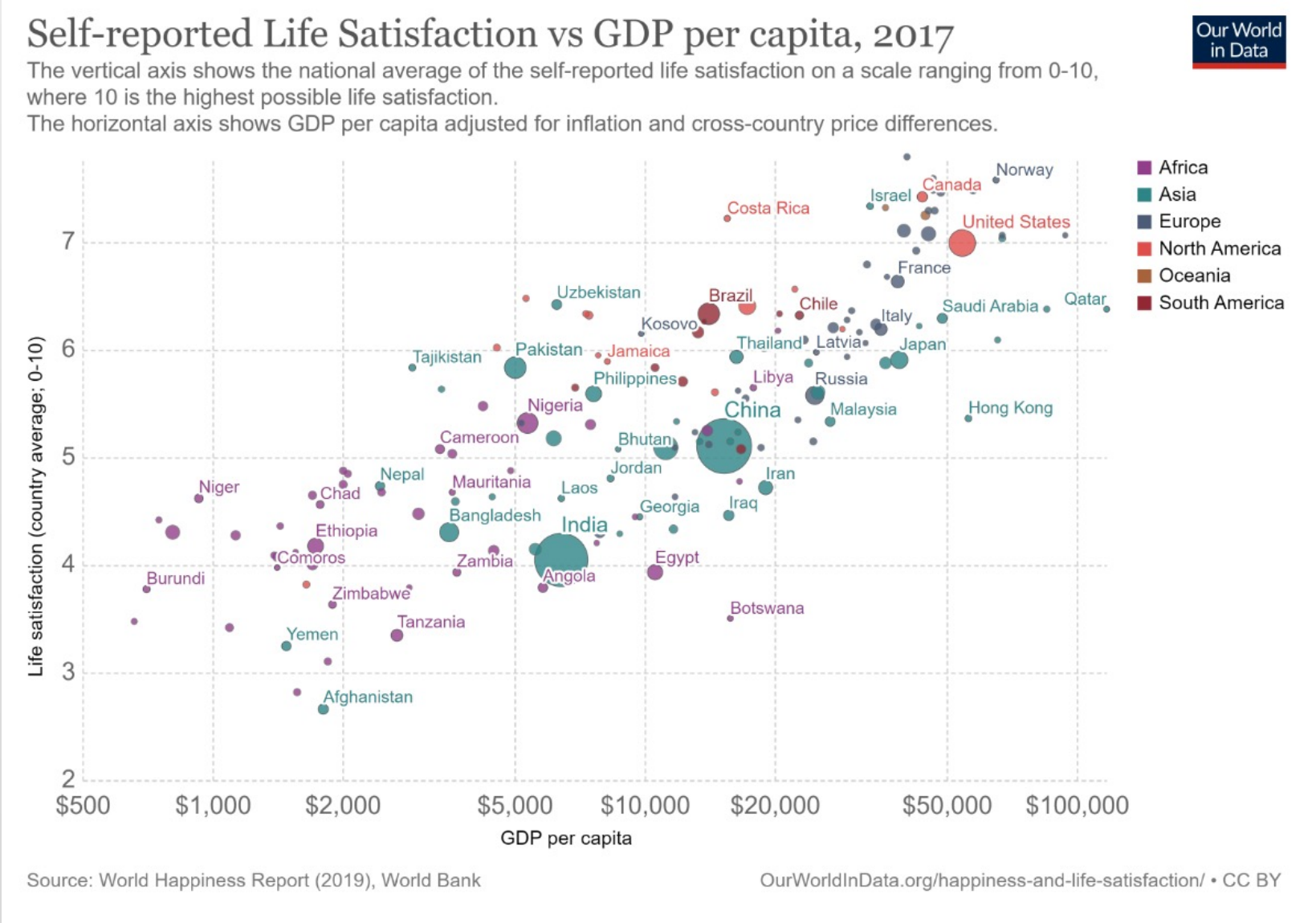
Corelația nu înseamnă cauzalitate

Uneori corelația nu înseamnă nimic

Covariația variabilelor

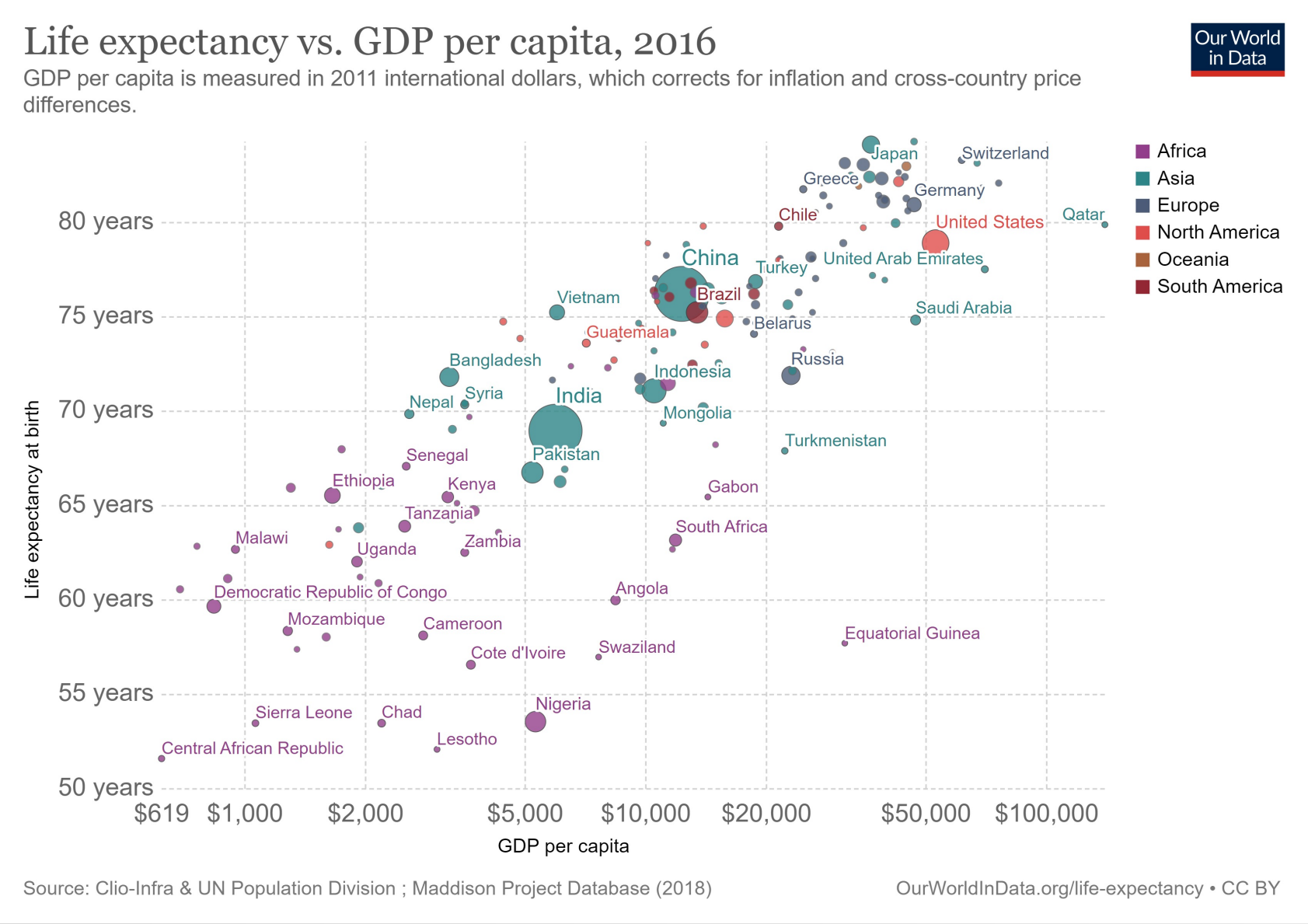
- Dacă ghicim mai bine valorile lui A știind valorile lui B decât la întâmplare, atunci spunem că:
  - A și B sunt asociate (vezi clarificări aici)
  - A și B covariază
  - A și B corelează (mai ales pentru variabilele numerice)
- Și variabilele ordinale sau nominale pot fi asociate
  - Județul și etnia
  - Țara și limba maternă

Corelație pozitivă



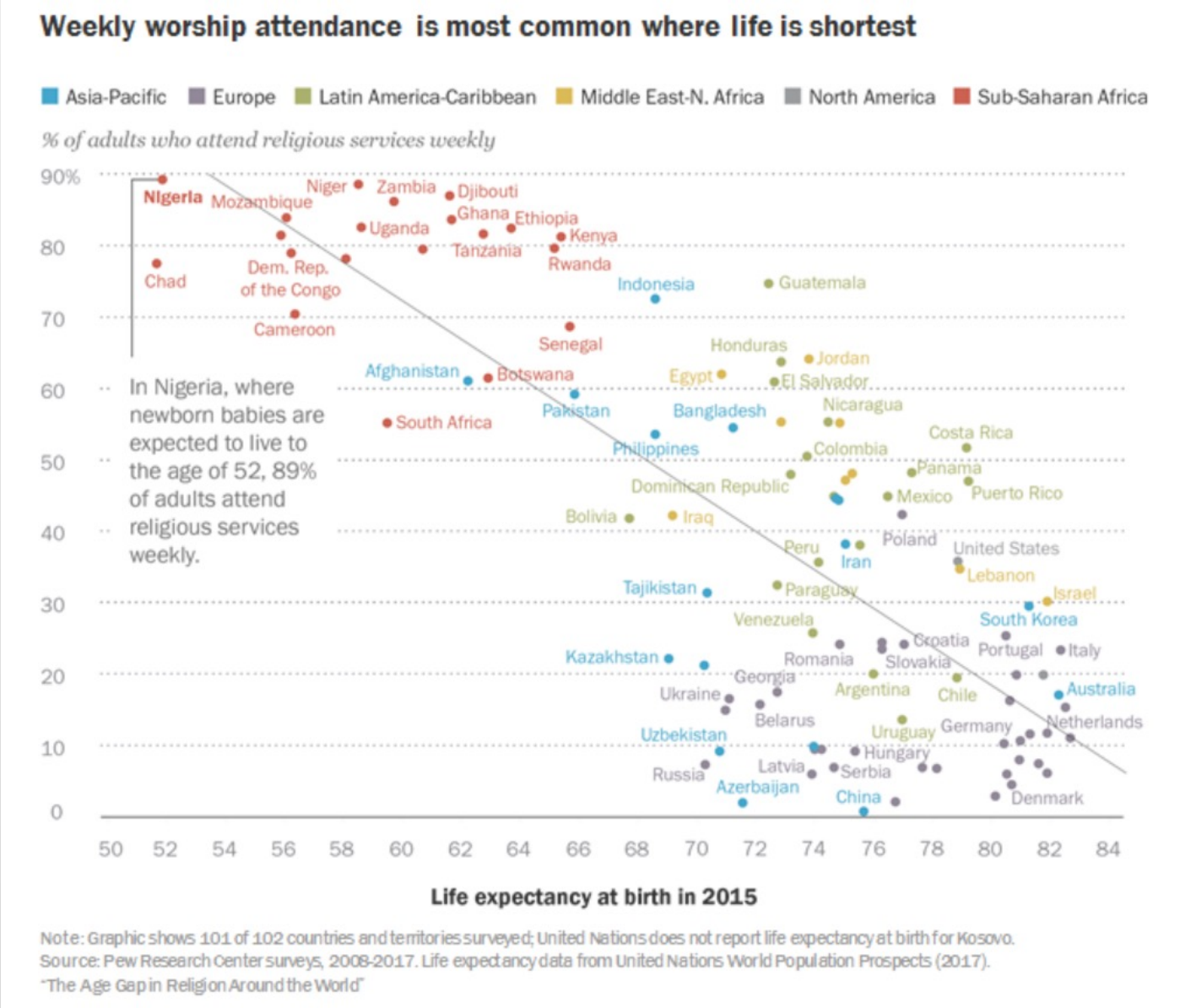
Sursă

Corelație pozitivă



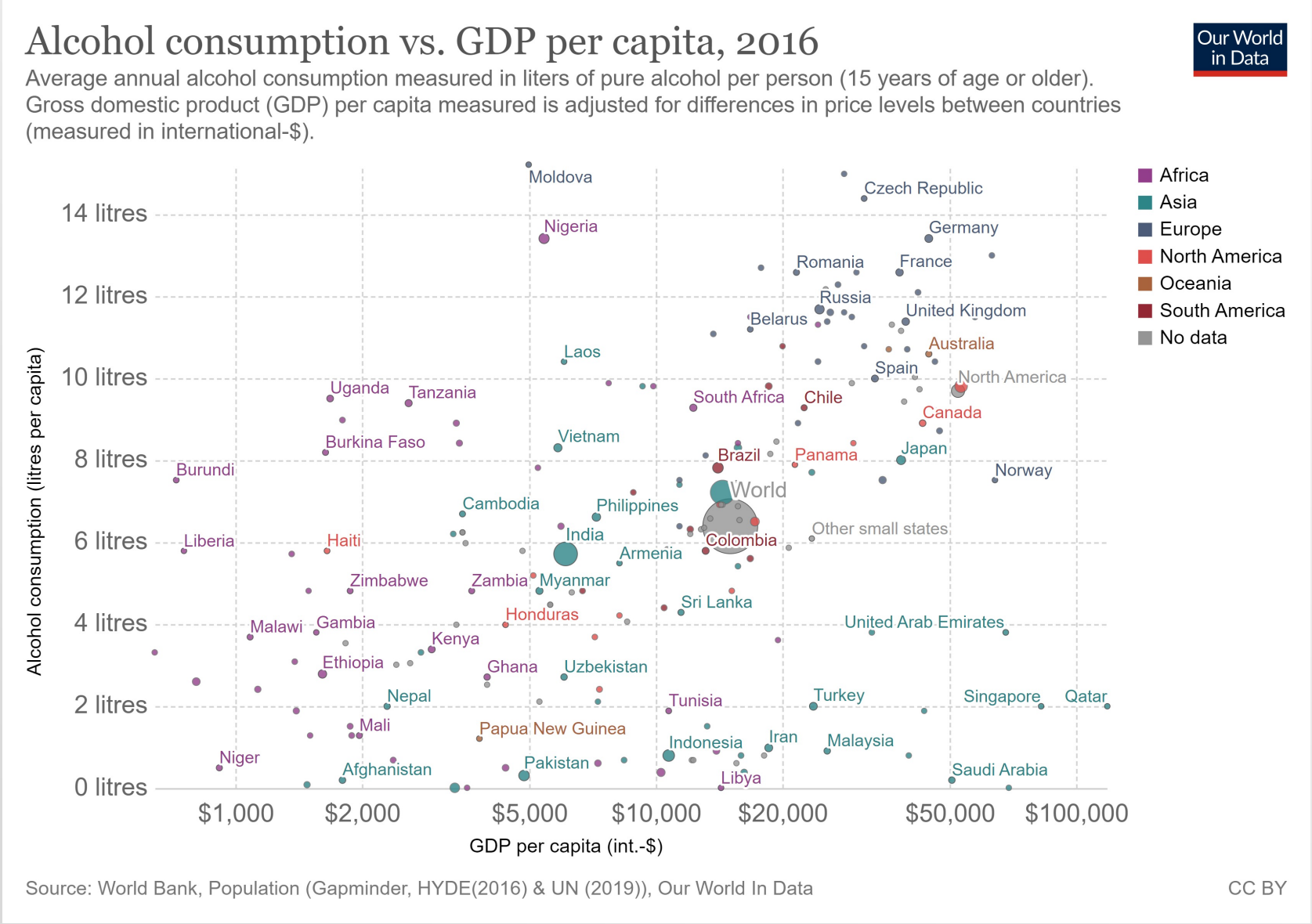
Sursă

Corelație negativă



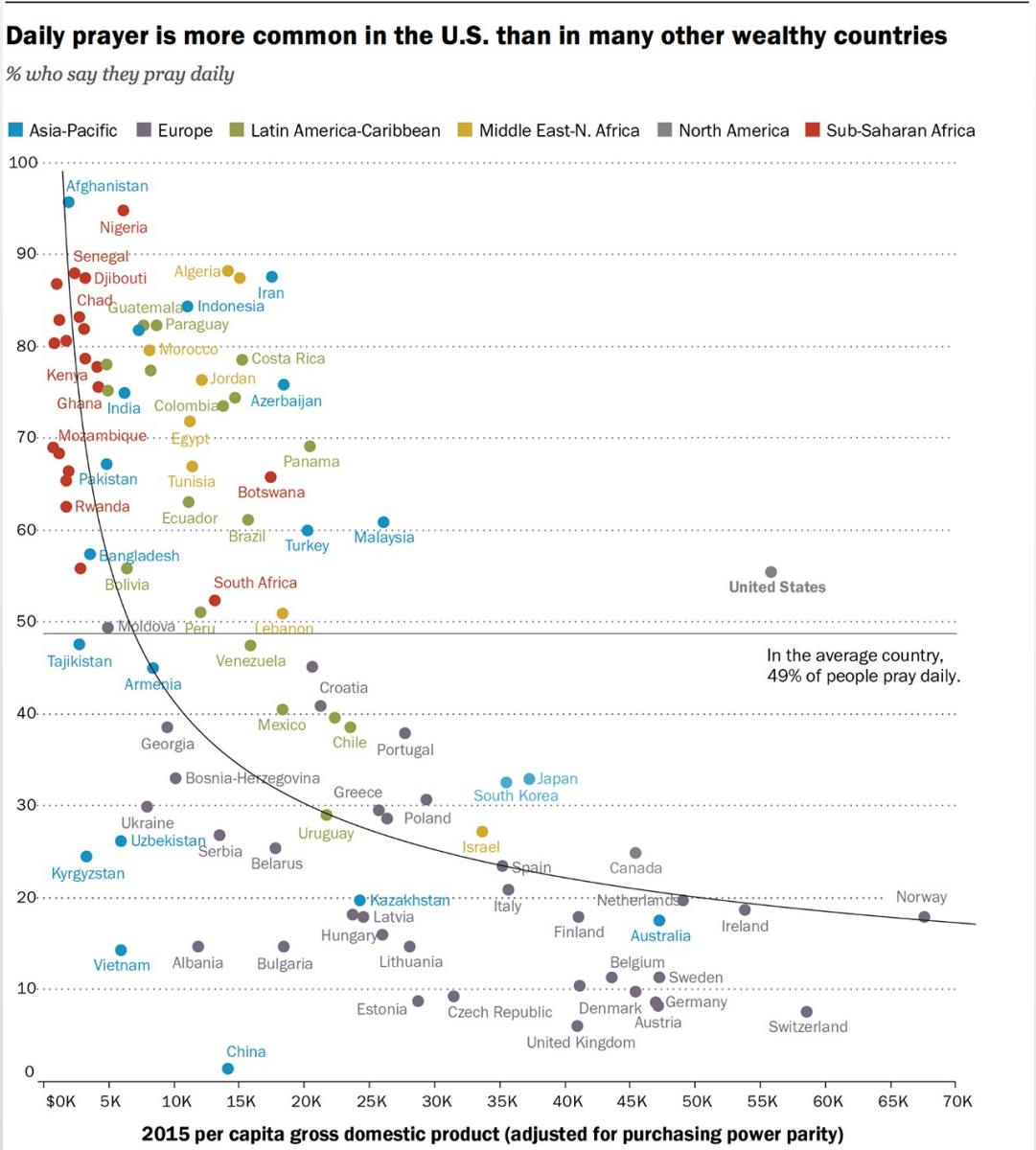
Sursă

Corelație nulă/slabă: noruri de puncte suprapuse



Sursă

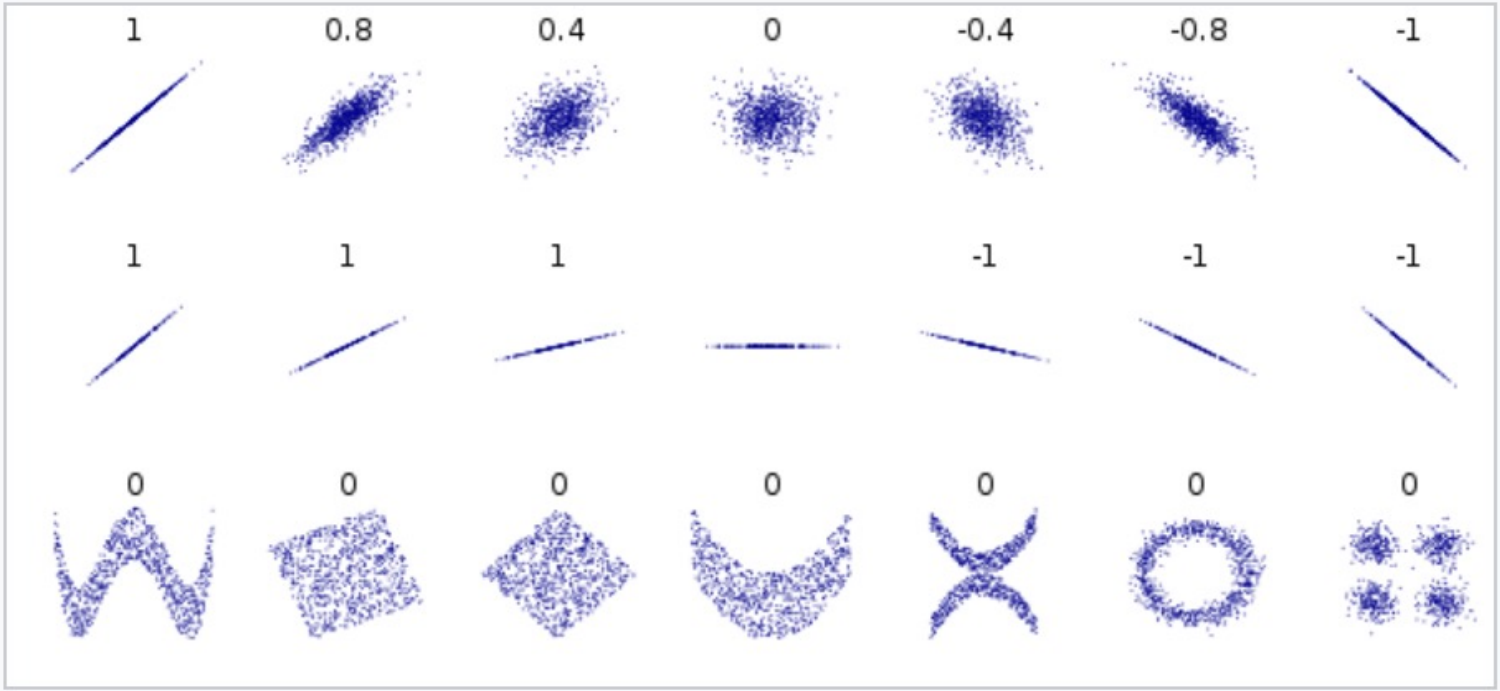
Corelație nonliniară



Sursă

Coeficientul de corelație R

- Dacă două variabile numerice co-variază liniar putem calcula coeficientul de corelație **Bravais-Pearson: R**
  - R măsoară intensitatea relației liniare dintre două variabile numerice
  - R variază între -1 și 1
  - Dacă R = 0 atunci variabilele nu sunt asociate
- Există și alte măsuri ale asocierii
  - Covarianța
  - Alți coeficienți speciali pentru variabile ordinale sau nominale



Several sets of (*x*, *y*) points, with the Pearson correlation coefficient of *x* and *y* for each set. The correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

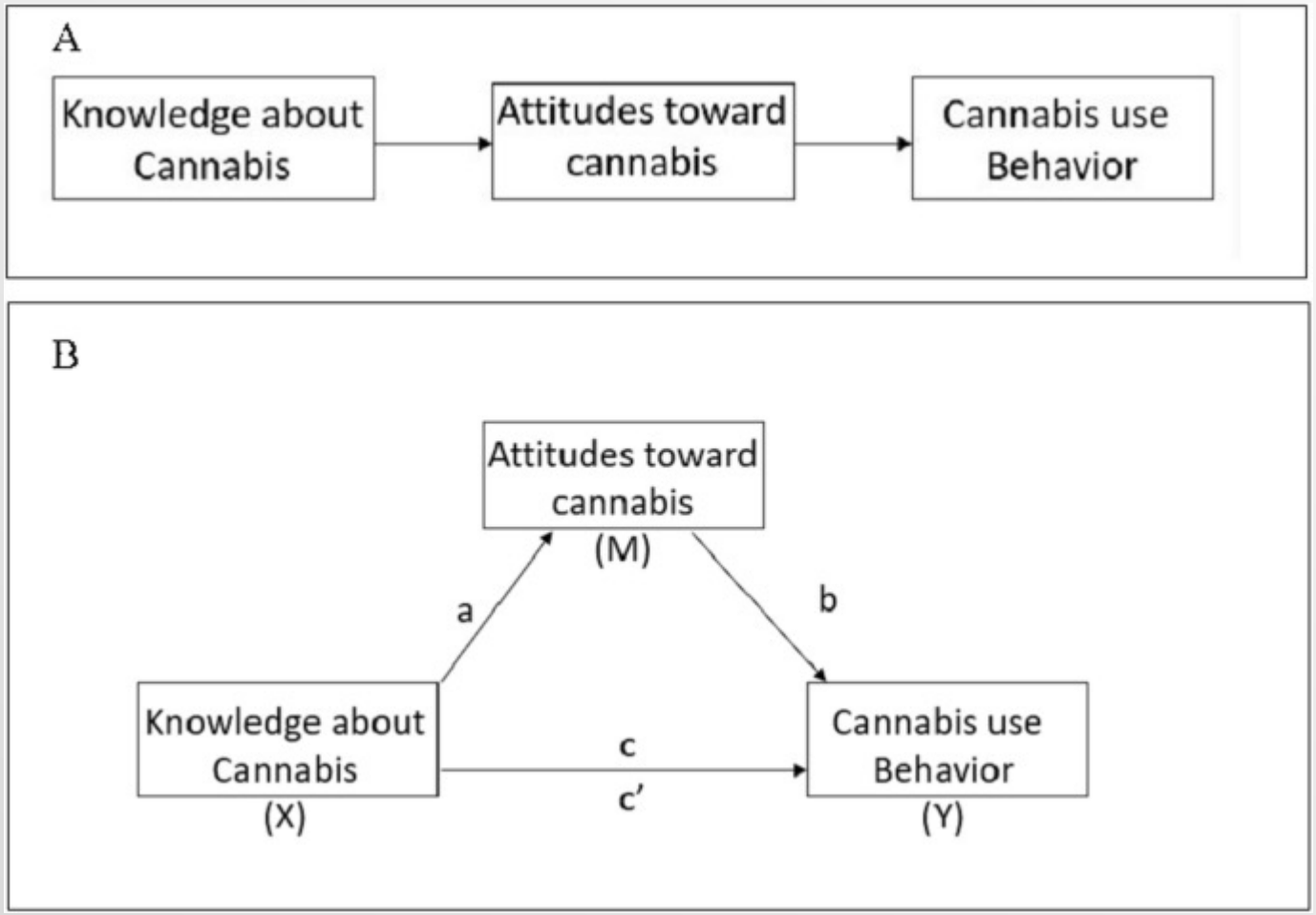
Corelații

- Care e coeficientul de corelație dintre...
  - Vârsta și anul nașterii unei persoane?
  - Numărul de căsătorii și numărul de divorțuri ale unei persoane?
  - Lungimea părului și mărimea pantofului?
  - Numărul de animale de companie avute și codul poștal?

#2.2 Erori de modelare

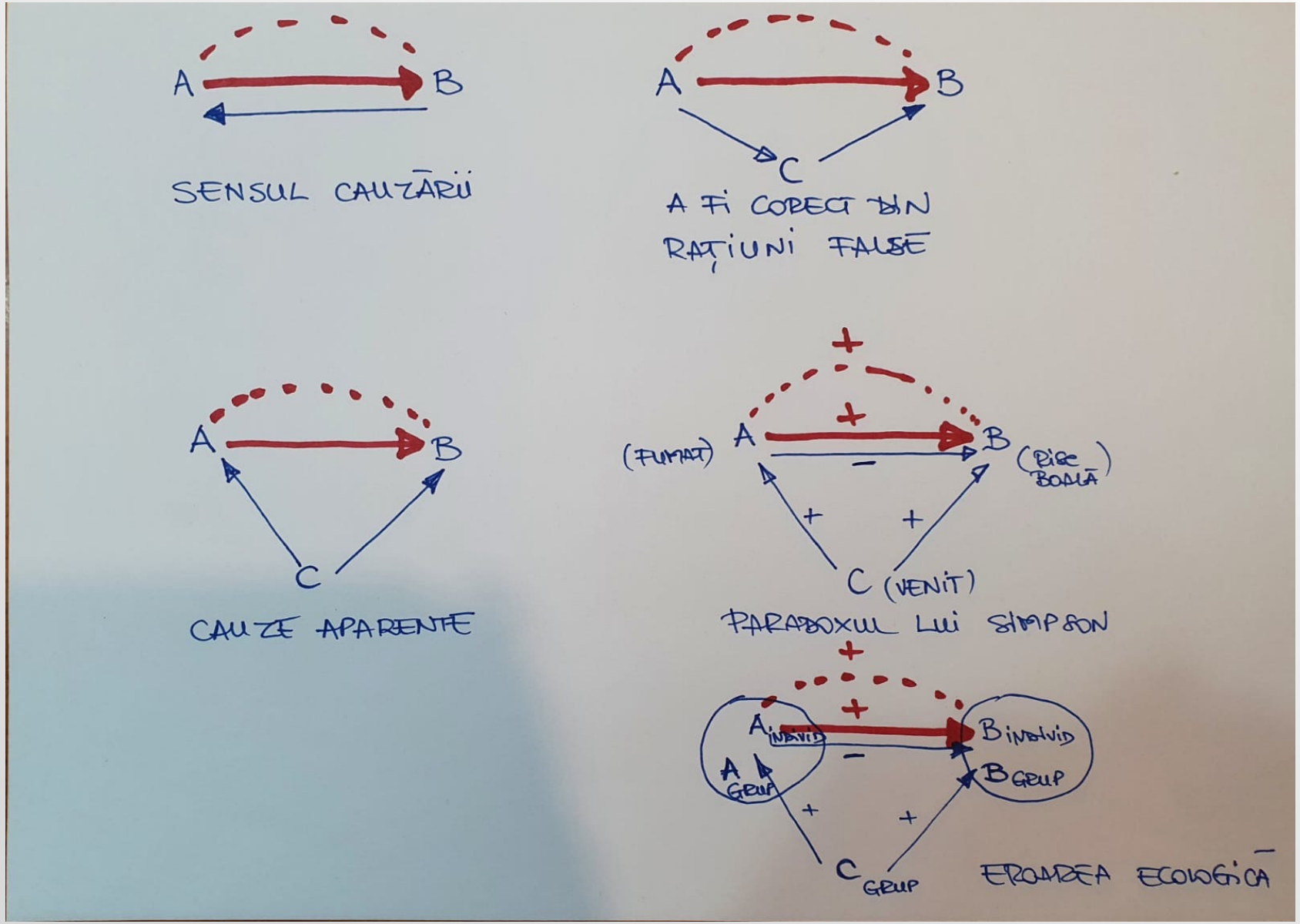
Modelarea

- Modelul face legătura dintre corelații și cauzalitate
- Aceleași corelații pot fi modelate diferit
  - Și greșit ☺
- Model = diagnoză



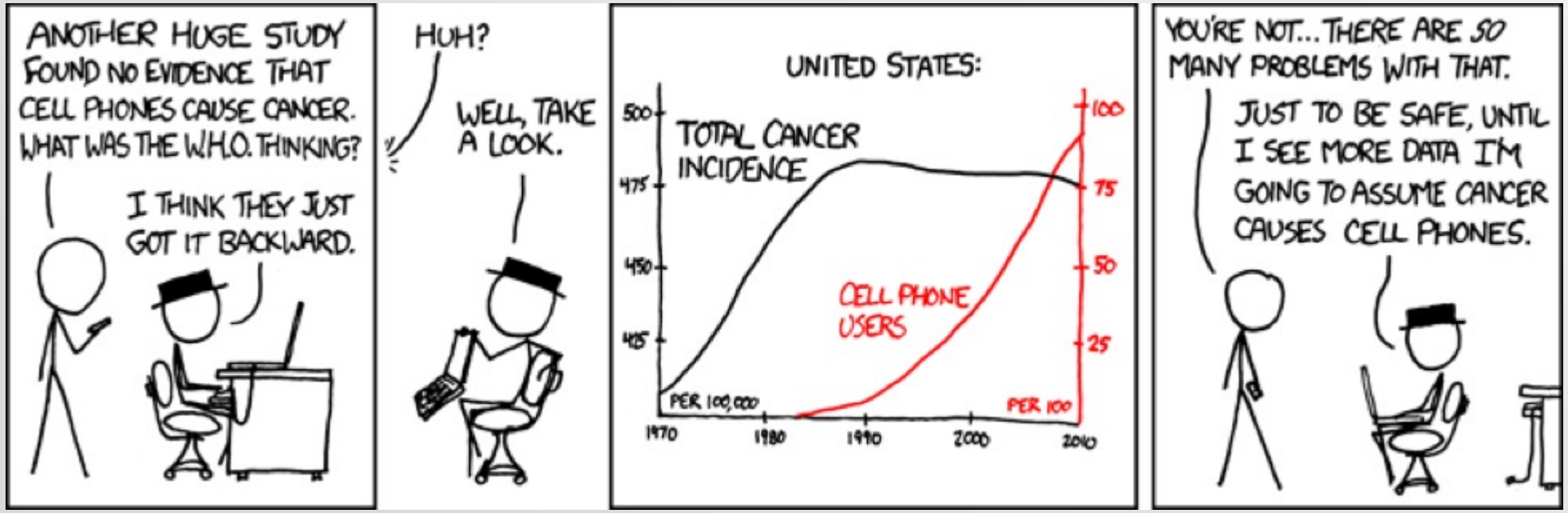
Zeiger et al.

Schemele grafice ale erorilor de modelare

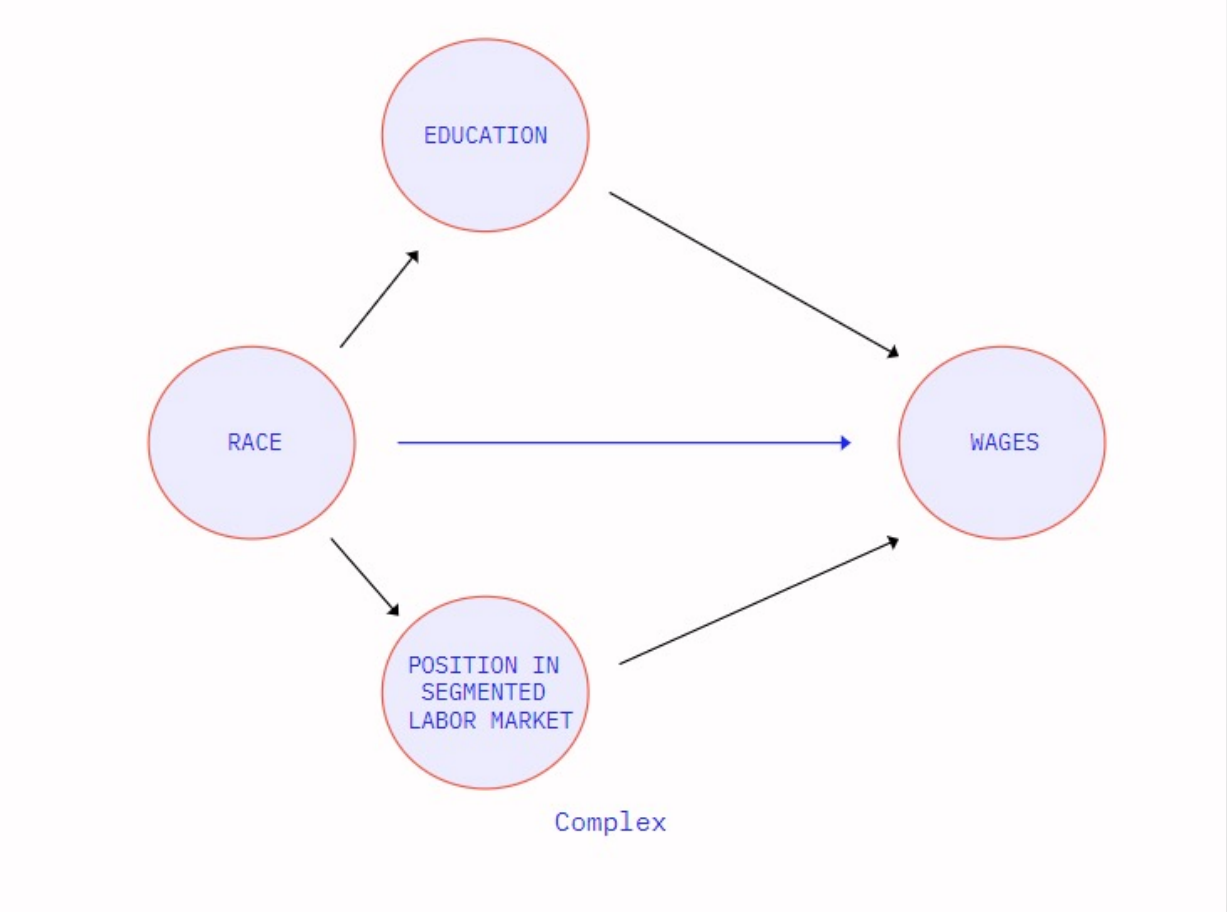


Sensul cauzării

- Care este relația dintre competență și săracie?
- Este sărăcia produsă de incompetența personală?
- Este sărăcia cauza incompetenței personale?

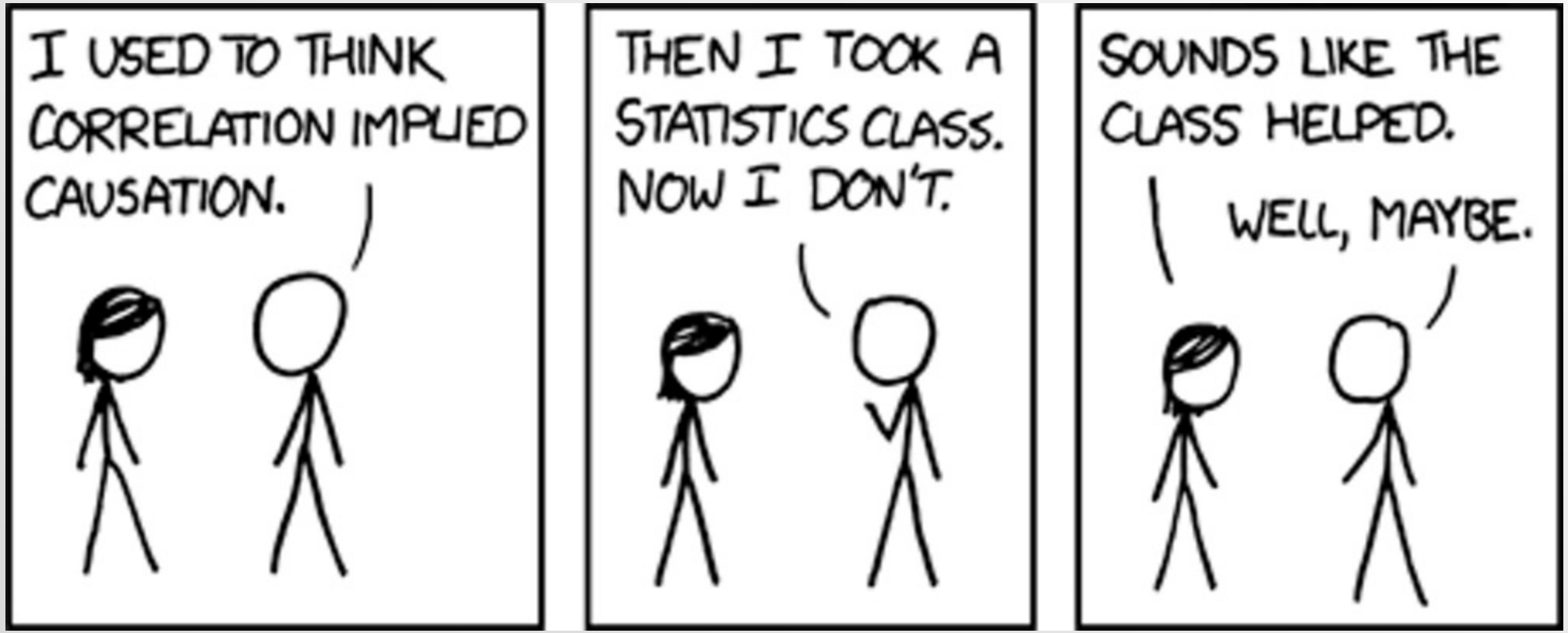


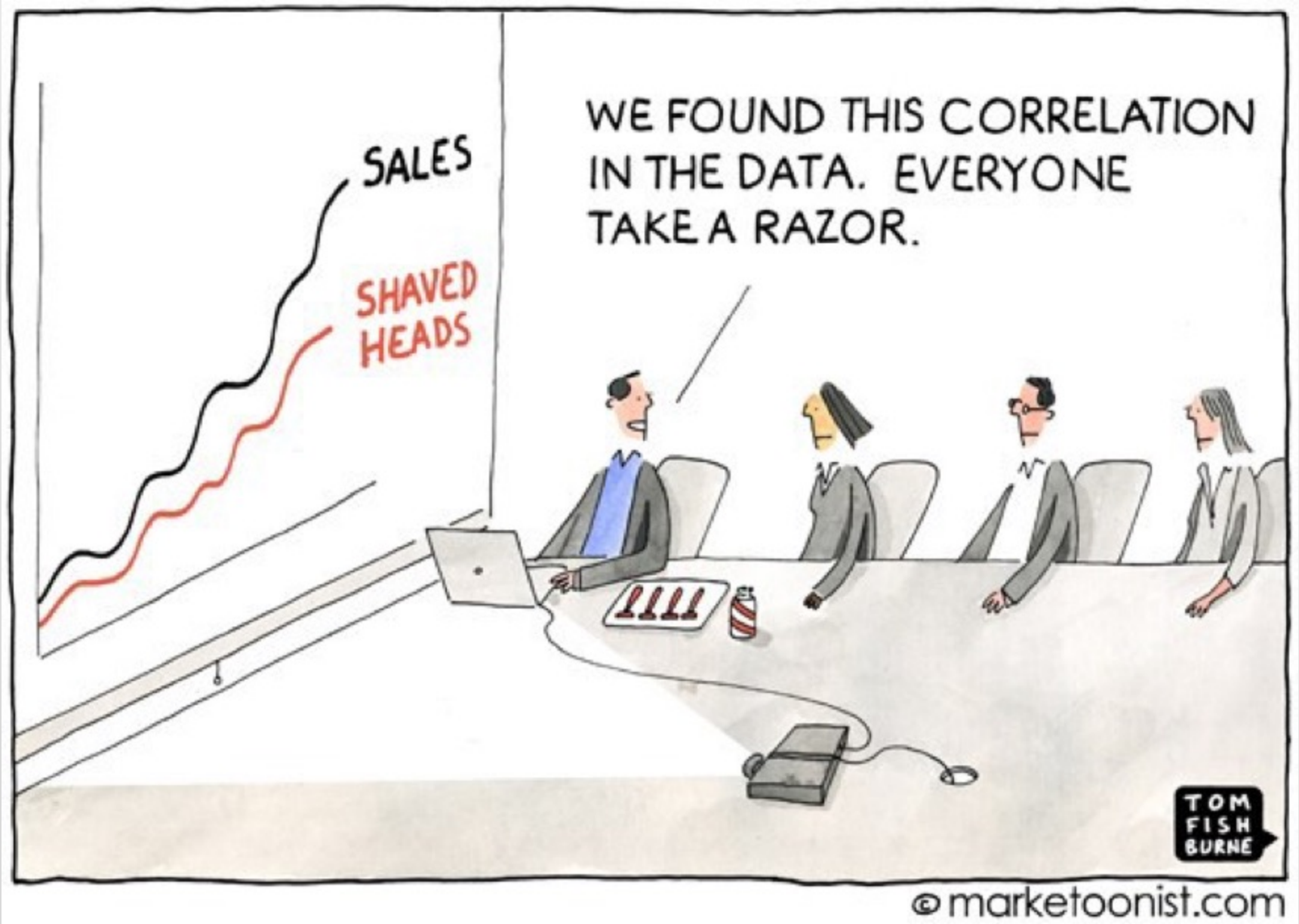
A fi corect din rațiuni false



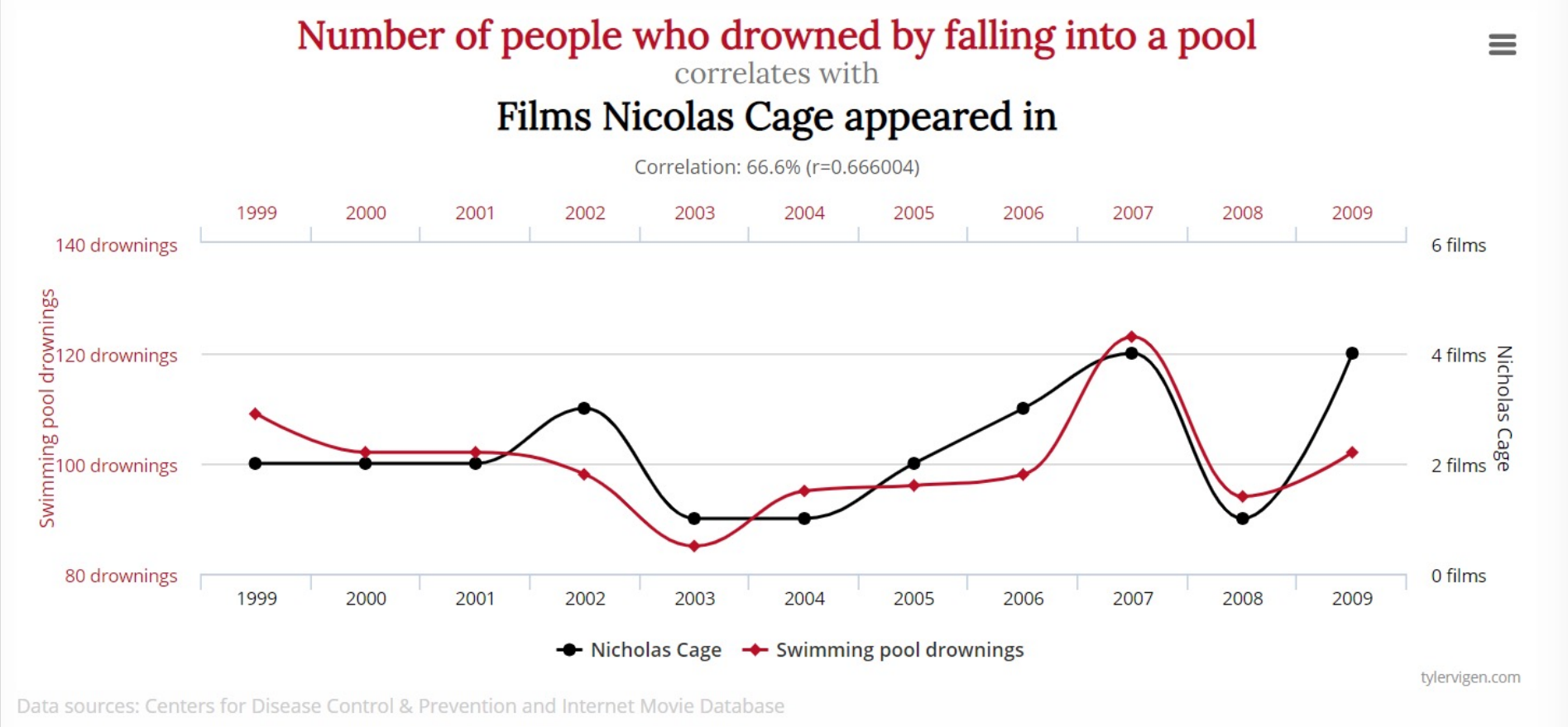
Corelații coincidentale

- Corelația nu înseamnă cauzalitate
- Uneori corelația nu înseamnă nimic
- Cu suficiente date putem găsi patternuri aleatoare ce par sistematice

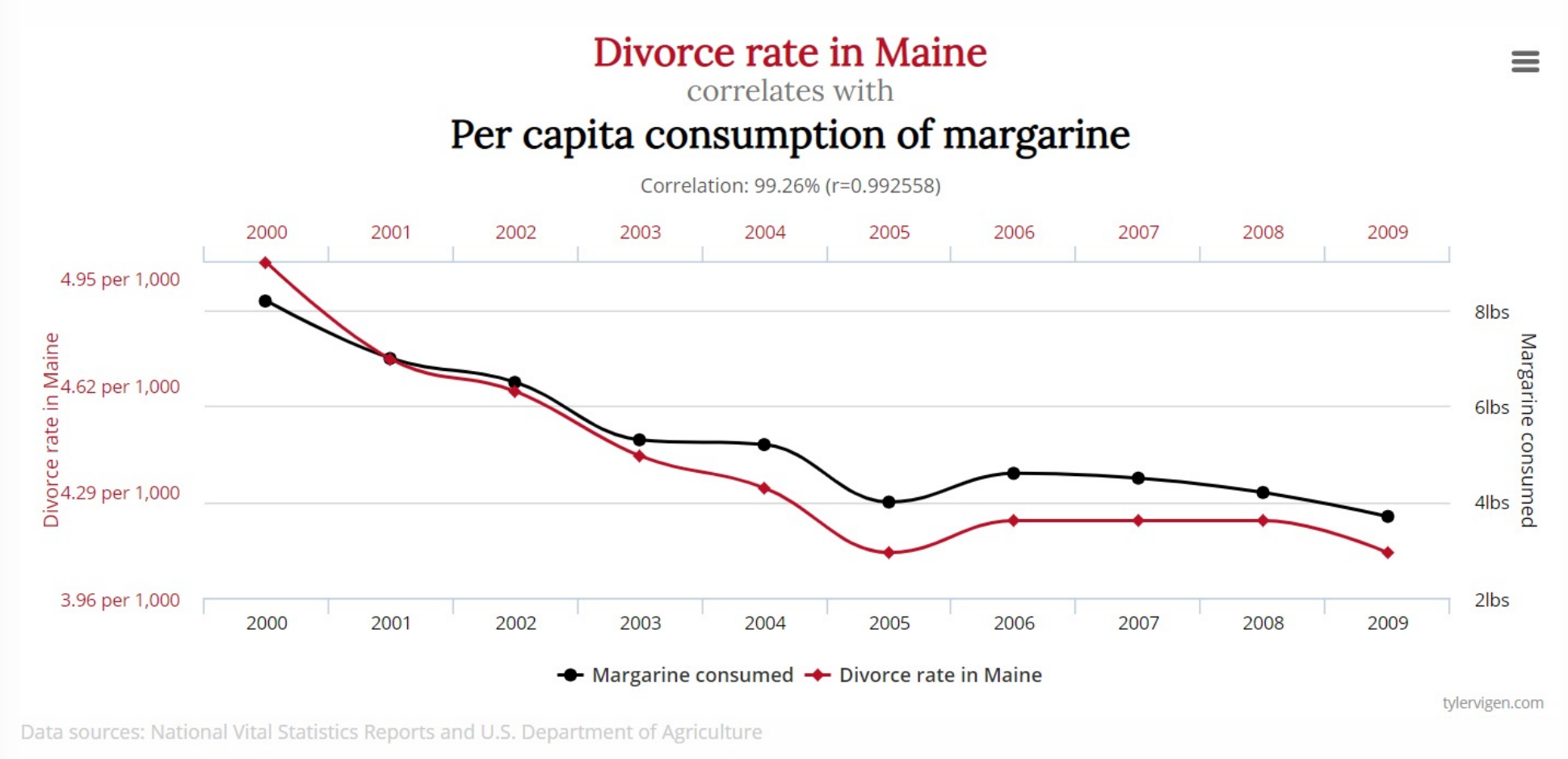

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.


SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISH BURNE
© marketoonist.com

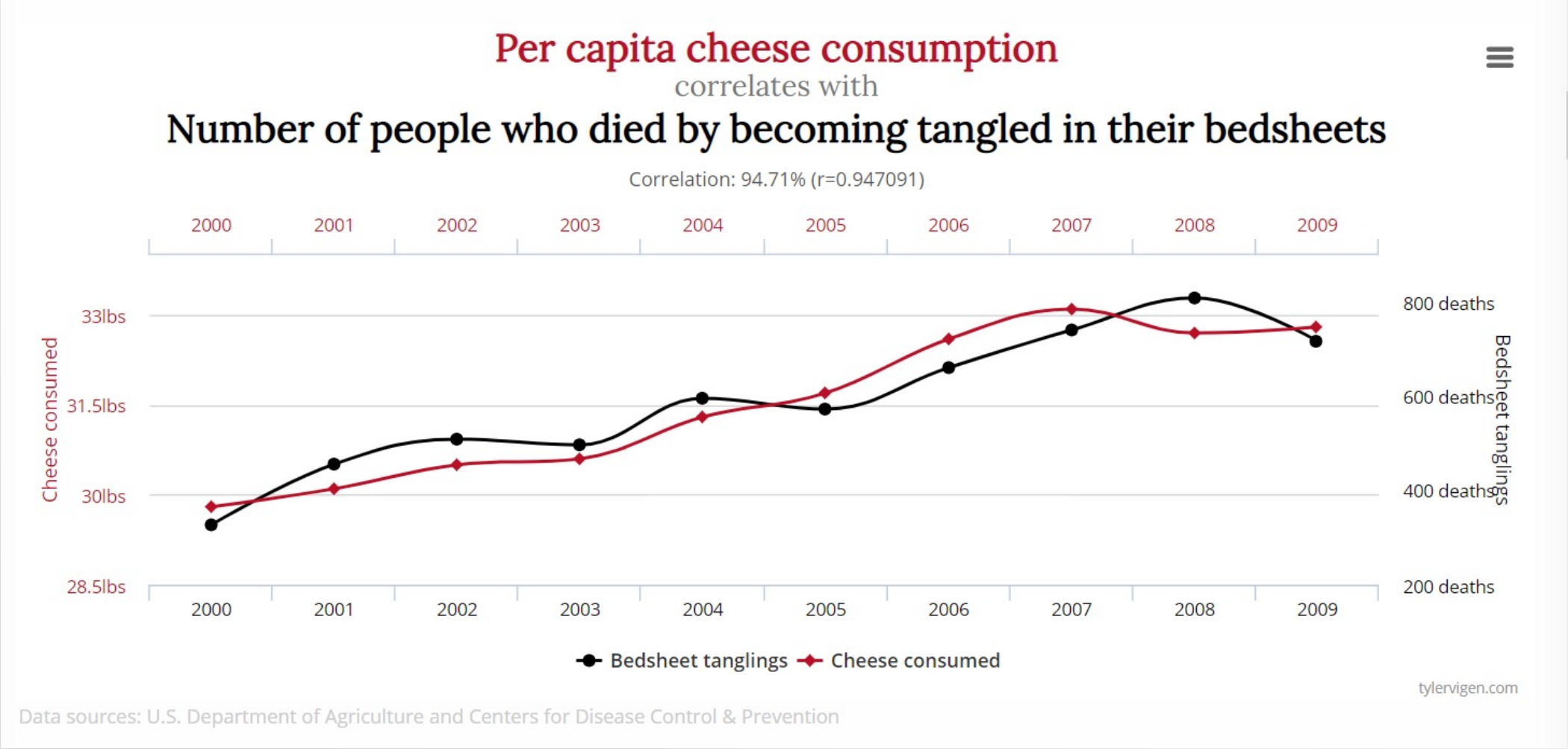
Corelații iluzorii - coincidentale



Tyler Vygen, Spurious correlations



Tyler Vygen, Spurious correlations

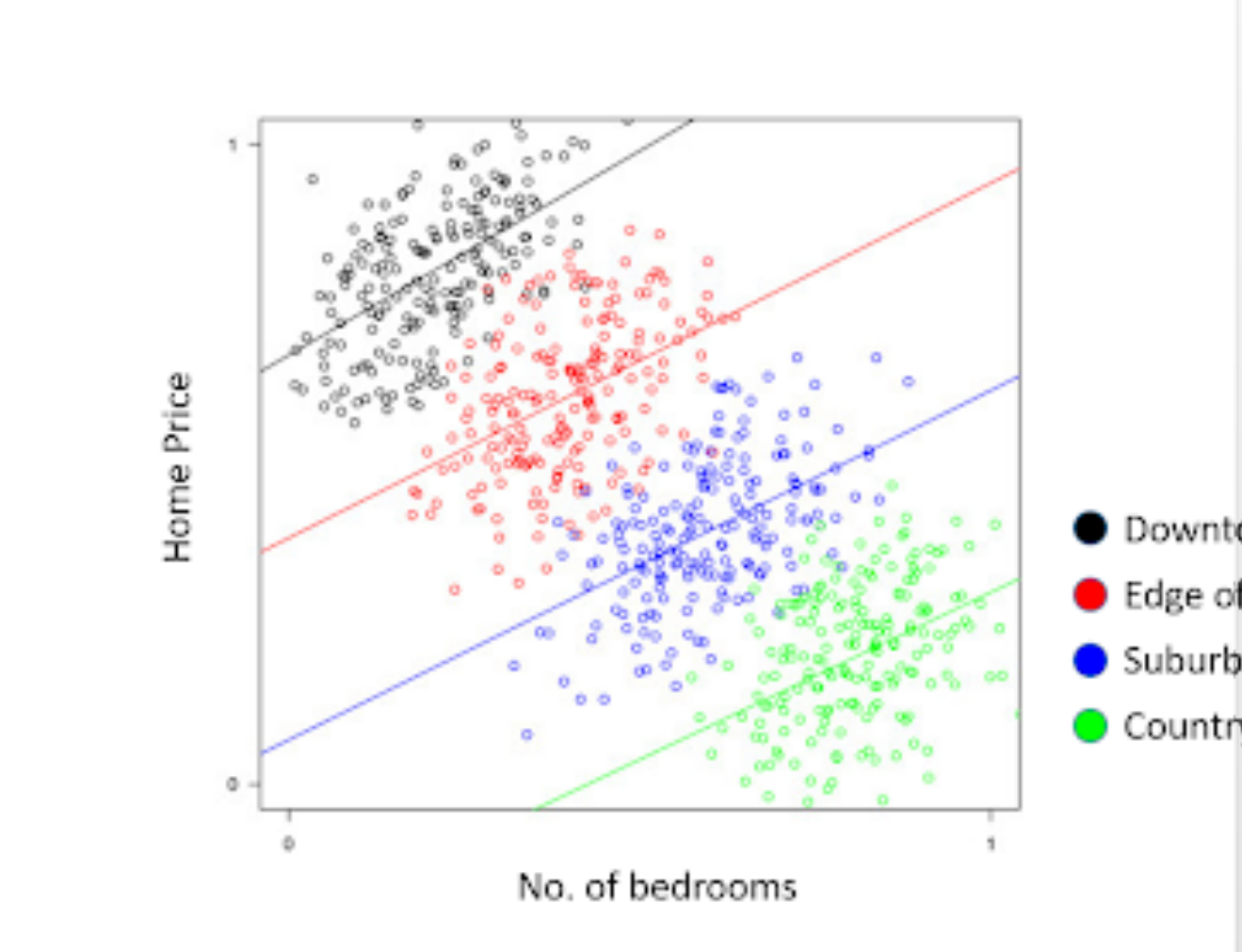


Tyler Vygen, Spurious correlations

Paradoxul lui Simpson

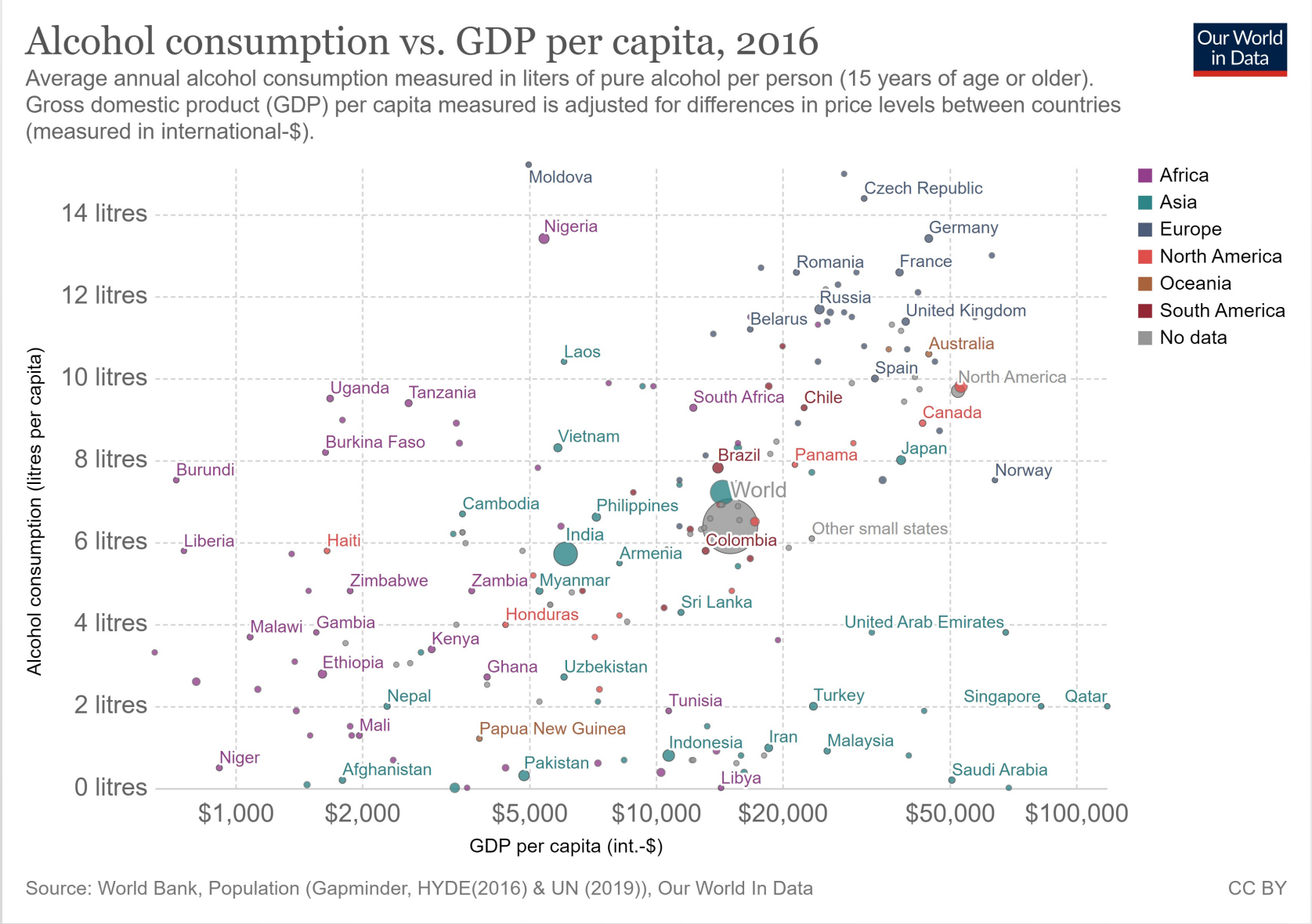
Modificarea unei relații atunci când ținem cont de o terță variabilă

(introducerea unei variabile de control)



[2]

Corelație slabă vs. pozitivă + nulă: controlul variabilei „religie”



Sursă

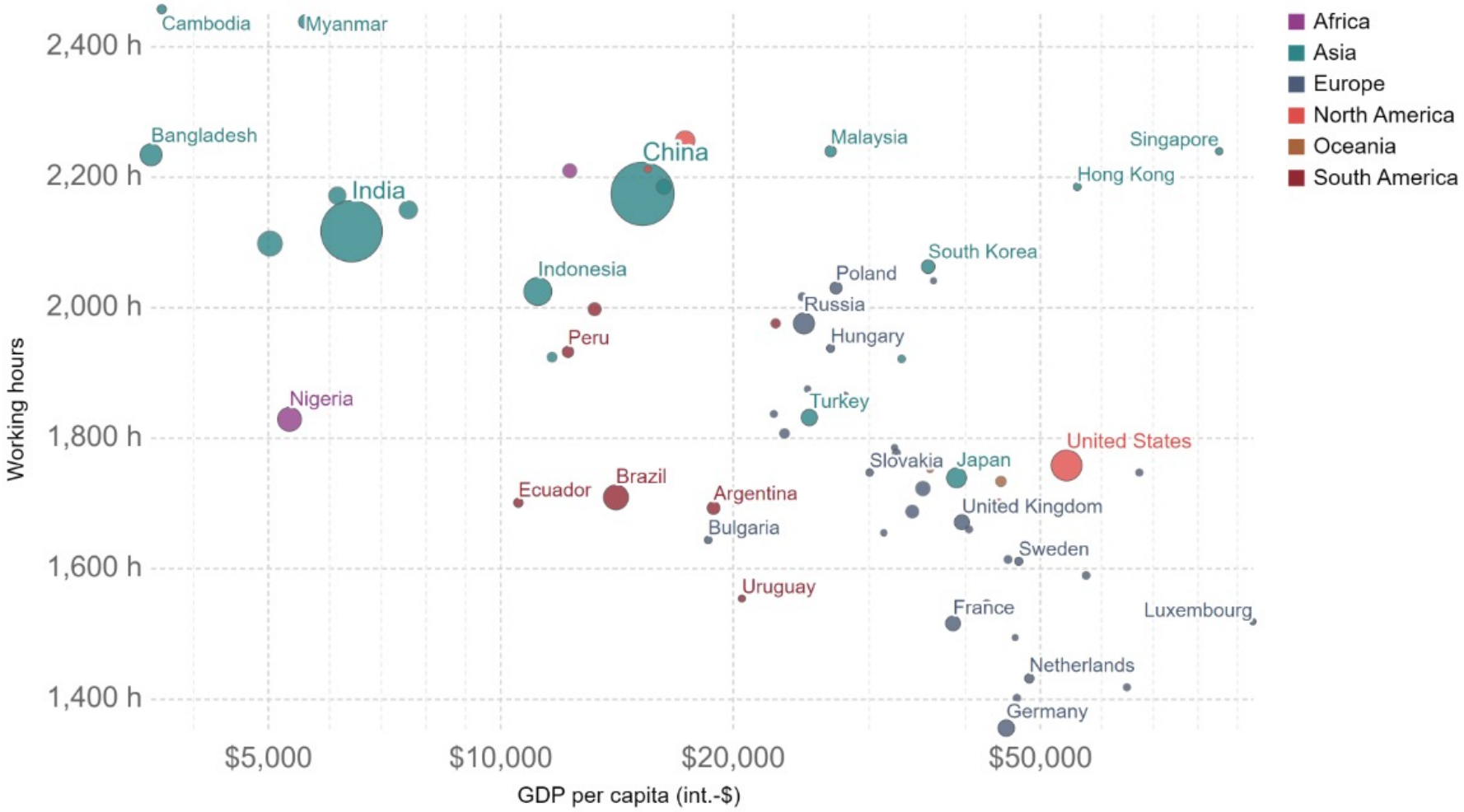
Eroarea ecologică

- Modificarea unei relații atunci când trecem de la unități individuale la unități agregate – sau invers
- Dacă în țările în care se muncește mult sărăcia e mare decât în celelalte, rezultă oare că în fiecare țară oamenii care muncesc mult sunt mai săraci decât ceilalți?
  - Relație la nivel agregat: unitatea = țara
  - Relație la nivel individual, în interiorul țărilor: unitatea = individul

Average annual working hours vs. GDP per capita, 2017

Working hours are the annual average per employed person. GDP per capita is measured in constant international-$. This means it is adjusted for price differences between countries and adjusted for inflation to allow comparisons between countries and over time.

Our World in Data



Source: Feenstra et al. (2015) Penn World Tables 9.1, World Bank

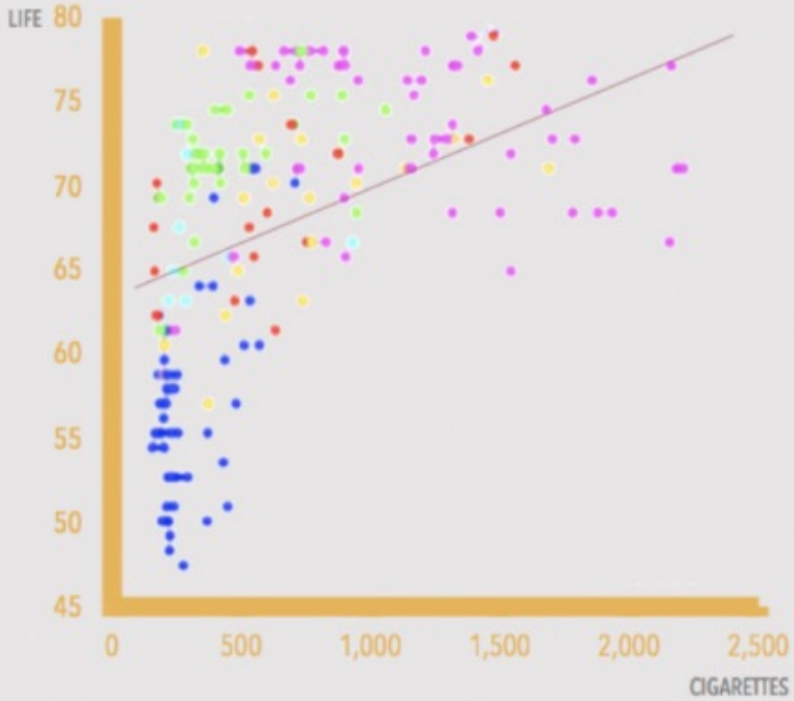
OurWorldInData.org/working-hours/ • CC BY

PR network crunch

Sursă

Eroarea ecologică / de agregare

Is smoking cigarettes good for your health?



Statistical Summary

Coefficients	Estimate	Std. Error	DF	t-value	p-value
(Intercept)	65.075840	0.855974	183	76.025515	<.00001
Cigarettes	0.006915	0.000855	183	8.090493	<.00001

This relationship is definitely statistically significant.

One extra cigarette per year adds 0.006915 years to your life.

So 4 cigarettes a day will add 10 years to your life.

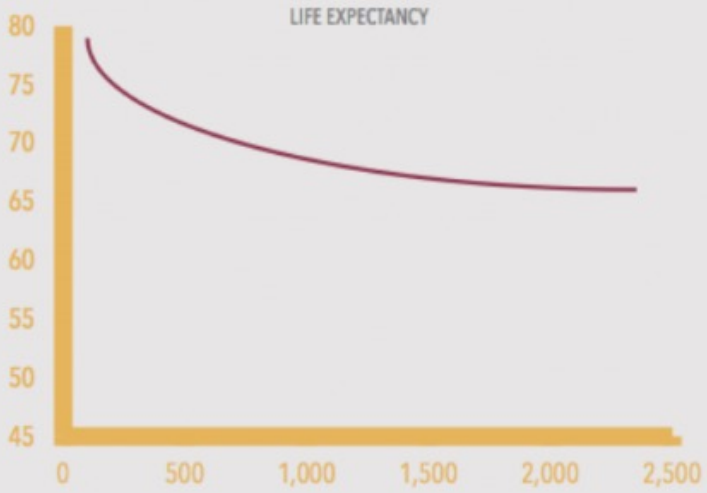
14

Krause, 2018

Eroarea ecologică / de agregare

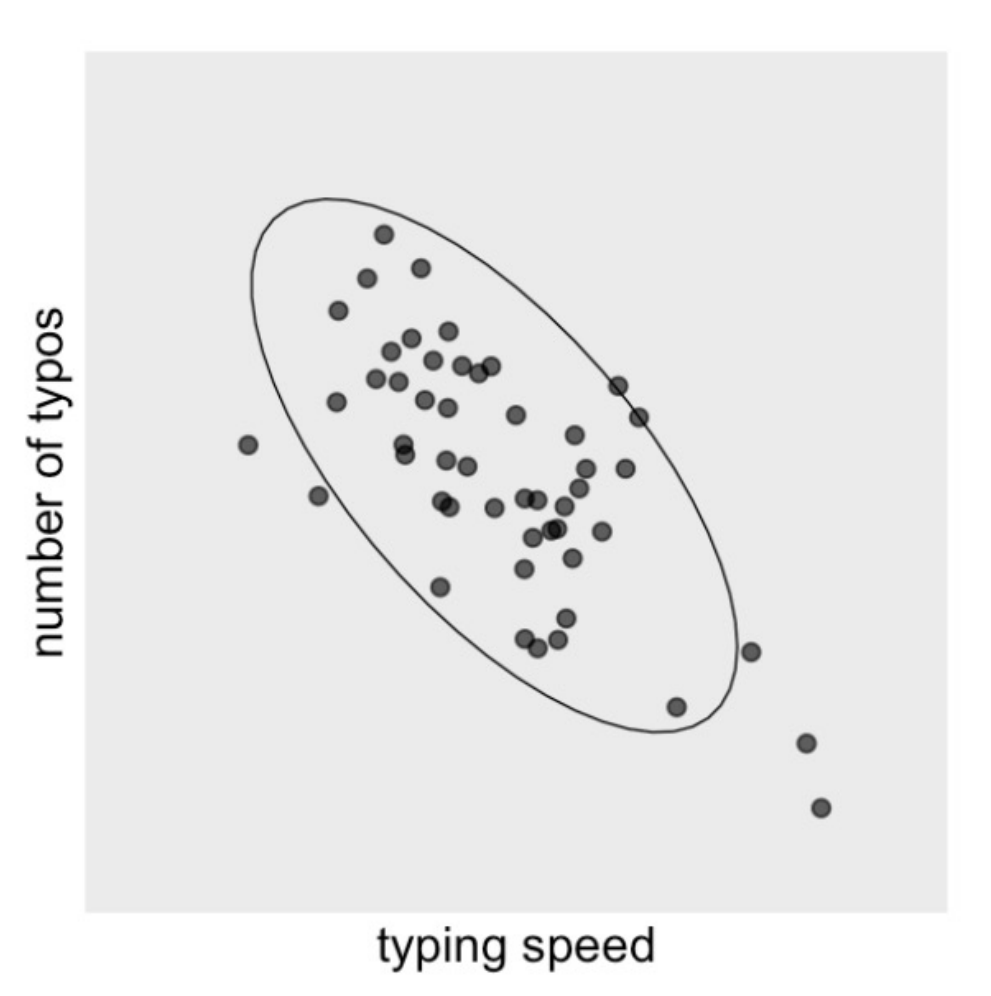
Is smoking cigarettes good for your health?

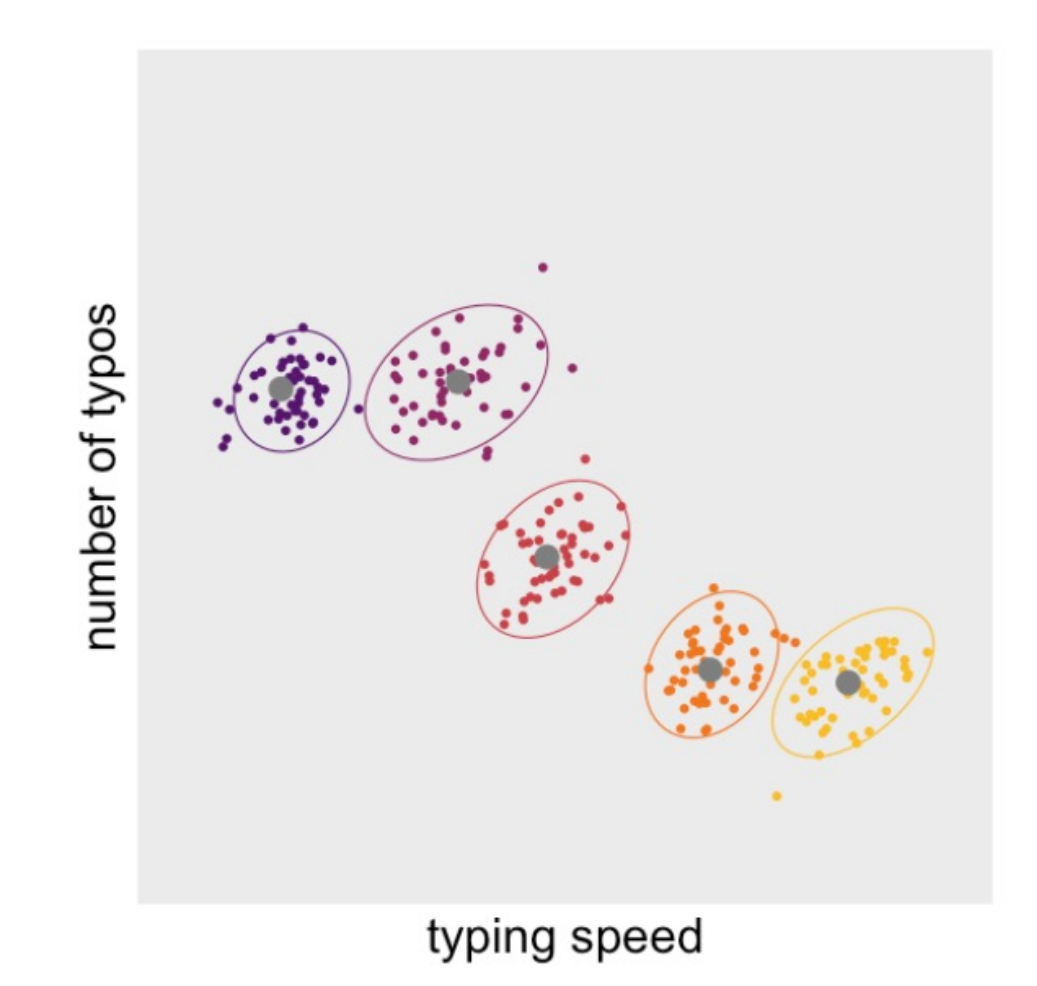
If we look at individual level data, we can see that the individual life expectancy decreases as smoking increases.



18

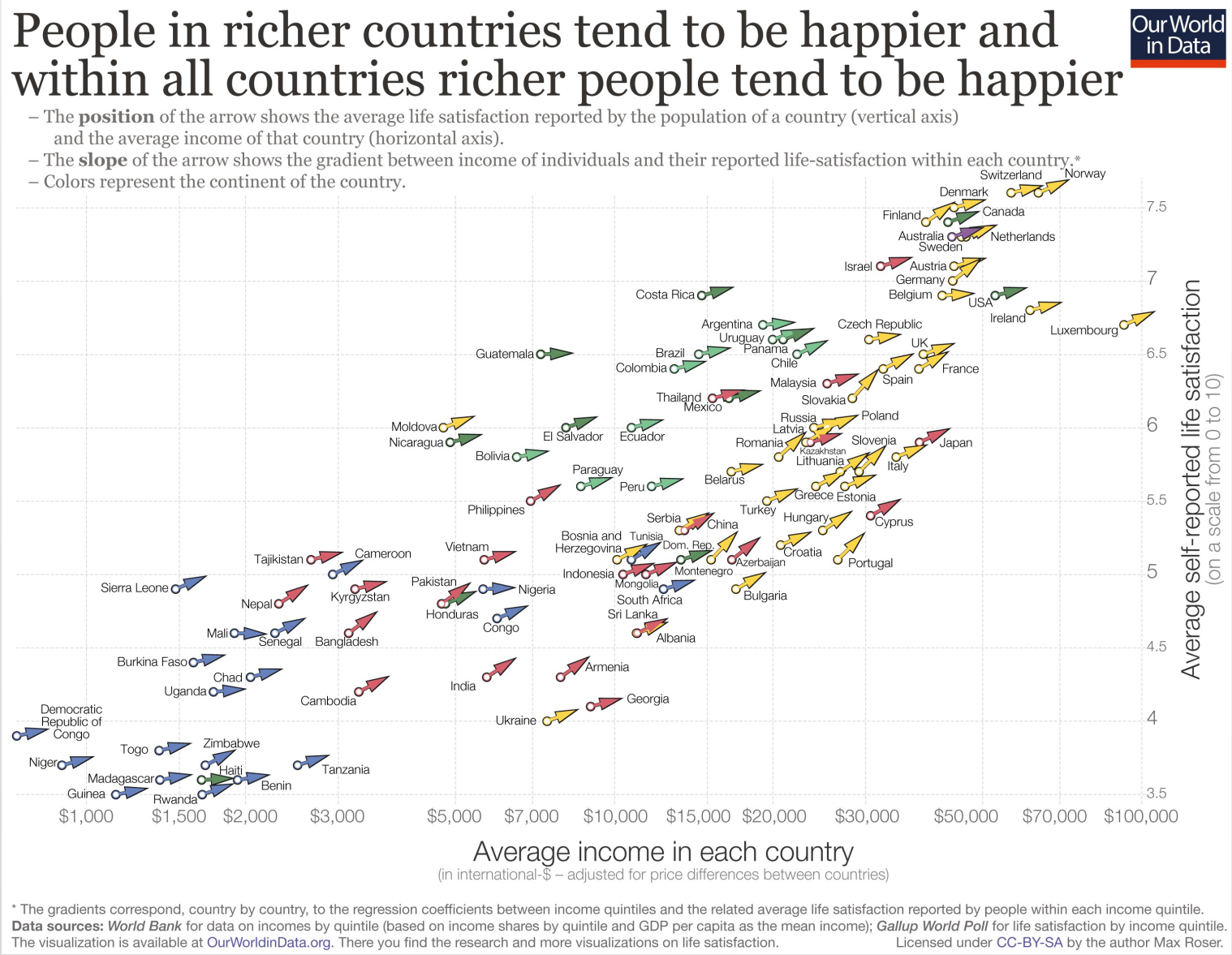
Eroarea ecologică / de agregare





SYSTEMS LABORATORY

Grafic care arată simultan relația la nivel agregat și la nivel individual



PR network crunch

Sursă

SYSTEMS LABORATORY

#2.3 Erori de eșantionare

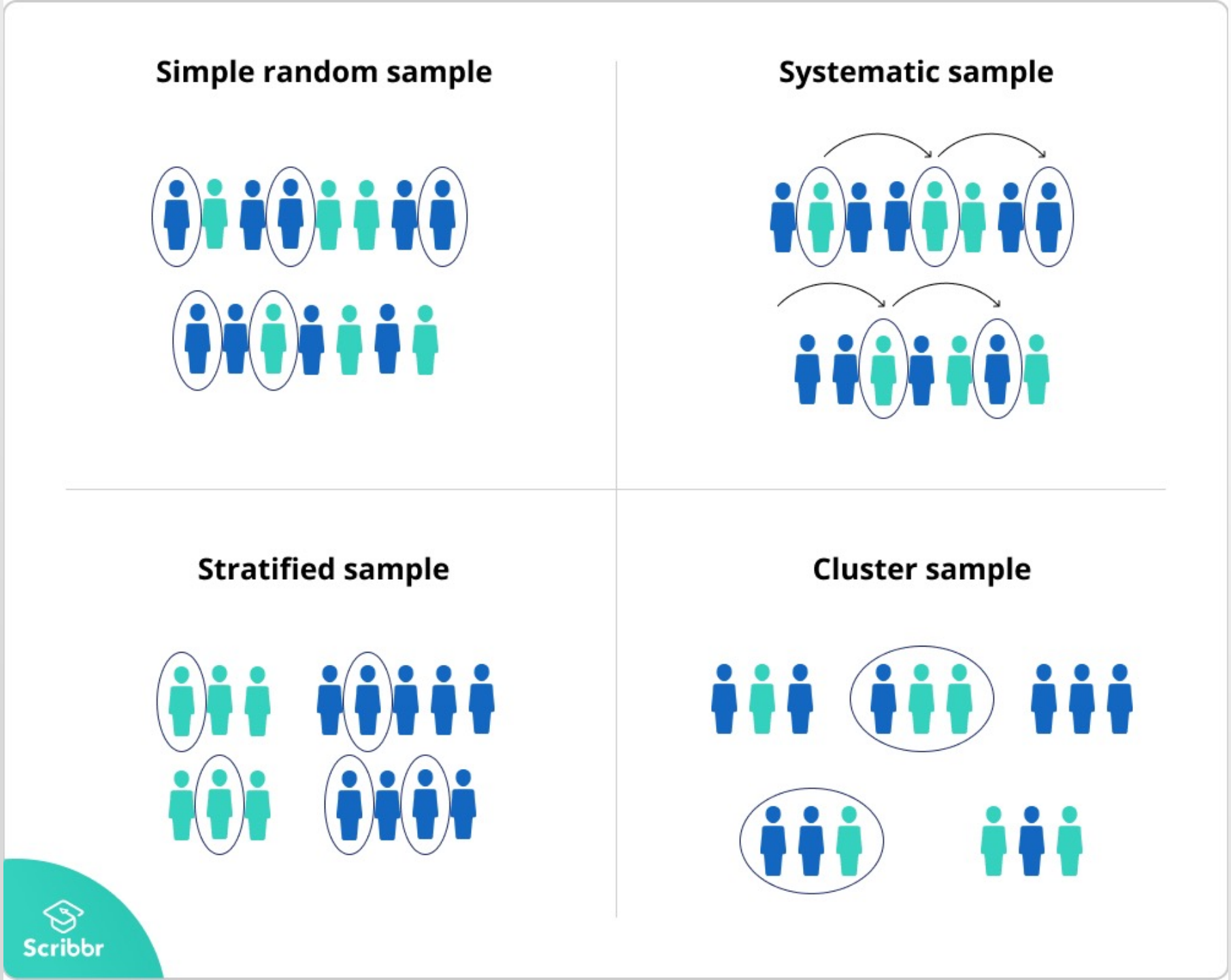
Eșantionare și generalizare

Probabilitatea de eroare (p)

PR network crunch

Eșantion și populație

- Nu putem studia toți membrii populației,
  - Studiem un eșantion
  - Generalizare (inducție) = tragem concluzii despre populație
- **Eșantionul probabilistic** permite **generalizarea**
  - Estimările obținute din eșantion sunt valabile, cu o marjă de eroare, pentru populație
- **Eșantionul non-probabilistic** nu permite generalizarea statistică
  - Eșantion de disponibilitate: comod dar ne-reprezentativ
  - Eșantion teoretic: permite explorarea sistematică a variației


Simple random sample
Systematic sample
Stratified sample
Cluster sample
Scribbr

[8]

Exemplu – Sondajele alegerilor SUA 2020

[What Went Wrong With Polling? Some Early Theories,](#) **NY Times**

The polls were off in 2020 in almost the same ways they were off in 2016.

	FINAL 2020 POLL AVG.	2020 POLLS WITH 2016 ERROR	2020 RESULT
U.S. †	+8 Biden	+6 Biden	+5 Biden
N.H.	+11 Biden	+7 Biden	+7 Biden
Wis.	+10 Biden	+4 Biden	<1 Biden
Minn.	+10 Biden	+4 Biden	+7 Biden
Mich.	+8 Biden	+4 Biden	+3 Biden
Nev.	+6 Biden	+8 Biden	+3 Biden
Pa. †	+5 Biden	<1 Biden	+1 Biden
Neb. 2*	+5 Biden	+9 Biden	+7 Biden
Maine 2*	+3 Biden	+9 Trump	+7 Trump
Ariz.	+3 Biden	+1 Biden	<1 Biden
Fla.	+2 Biden	<1 Biden	+3 Trump
N.C.	+2 Biden	+3 Trump	+1 Trump
Ga.	+2 Biden	<1 Biden	<1 Biden
Ohio	<1 Trump	+6 Trump	+8 Trump
Iowa	+1 Trump	+5 Trump	+8 Trump
Texas	+2 Trump	+4 Trump	+6 Trump

Cei care răspund sunt...
- mai agreabili (OCEAN)
- mai încrezători
- mai liberali (Vox)

One pollster's explanation for why the polls got it wrong

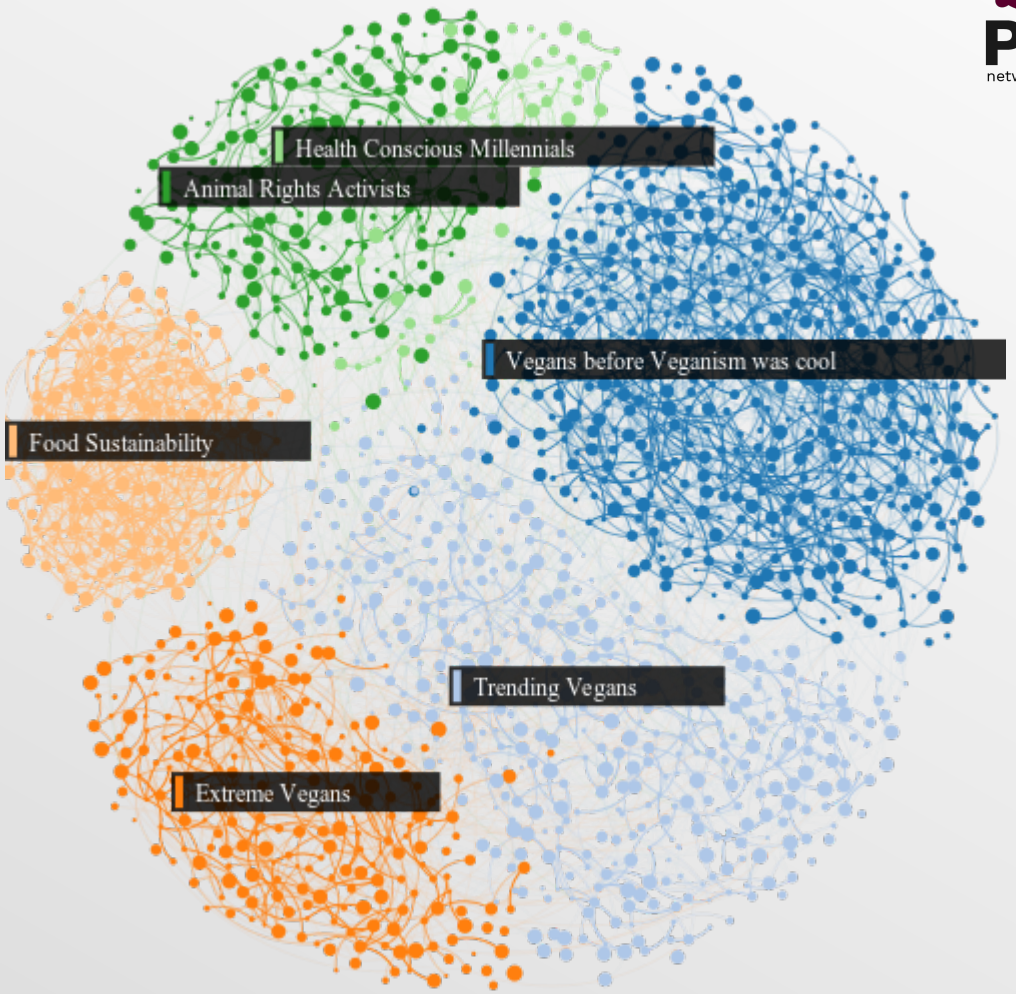
The kind of people who answer polls are really weird, and it's ruining polling.

By Dylan Matthews | dylan@vox.com | Nov 10, 2020, 9:20am EST

SHARE

Exemplu

- Sunt utilizatorii Twitter reprezentativi pentru populație?
  - „The Affinio algorithm ingests massive amounts of Twitter audience data – or social signals – and uncovers naturally-occurring actionable insights derived from a significant, unbiased, representative audience.” (Affinio)



(Davies 2019)

Semnificația statistică (p)

- Dacă avem **estimări bazate pe selecții probabiliste**, putem calcula p
  - Eșantioane aleatoare ale unei populații
  - Comparații între loturi cu alocare aleatoare
- p = „semnificația statistică”
- p = „probabilitatea de eroare”
- p = probabilitatea ca valoarea obținută în eșantion să fie datorată șansei și nu unui proces real în populație
- De regulă daca **p<5%** considerăm că acea valoare este reală în populație
- **Valoarea p nu are sens dacă nu se referă la selecții probabiliste!**
- **Valoarea p nu exclude erorile de modelare!**

[4]

Hypothesis:

Drinking rosé leads to shrinkage

[4]

SYSTEMS LABORATORY

#2.4 Erori de modelare – Studiu de caz

PR network crunch

F. Messerli, Consumul de ciocolată și premiile Nobel

The NEW ENGLAND JOURNAL of MEDICINE

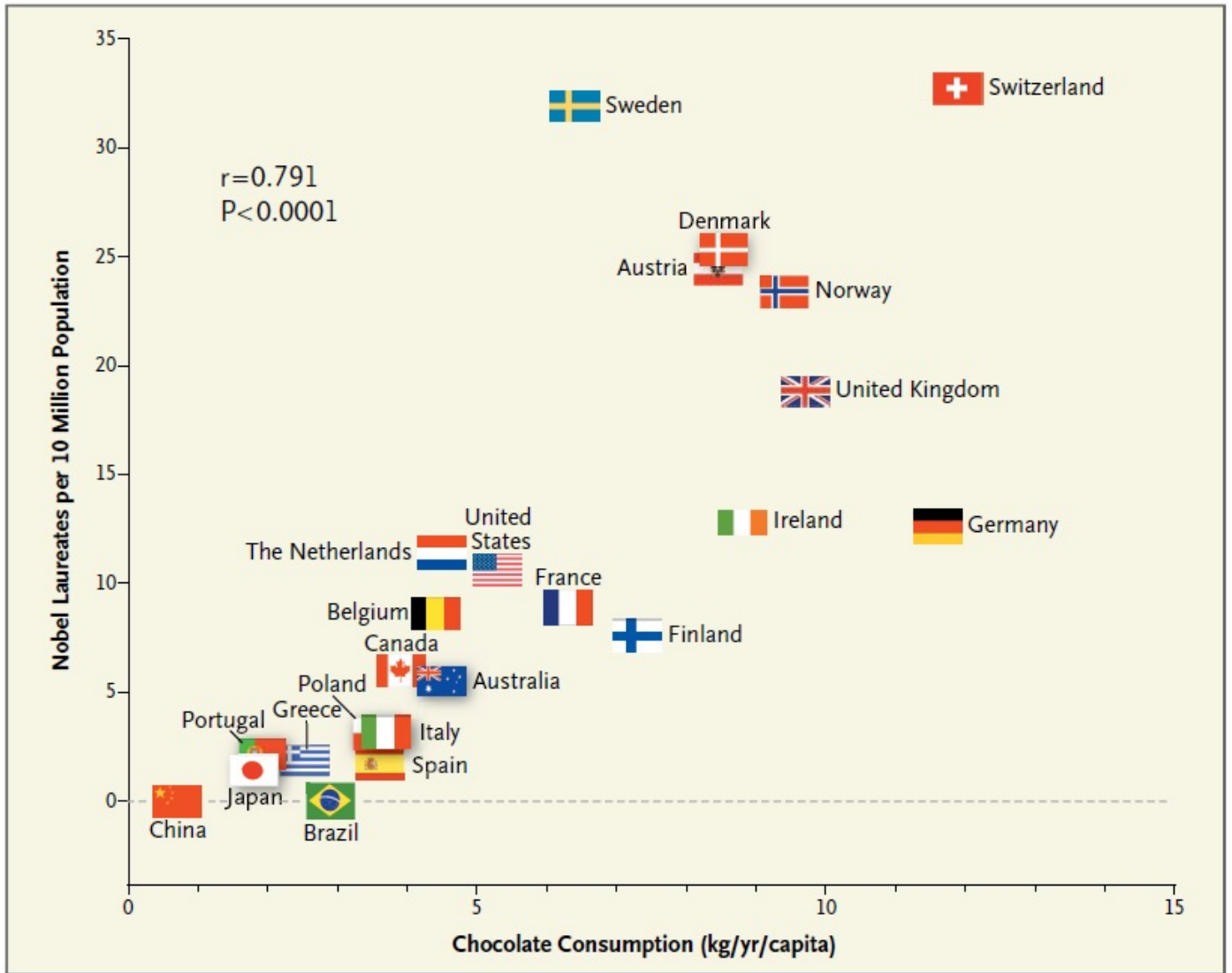
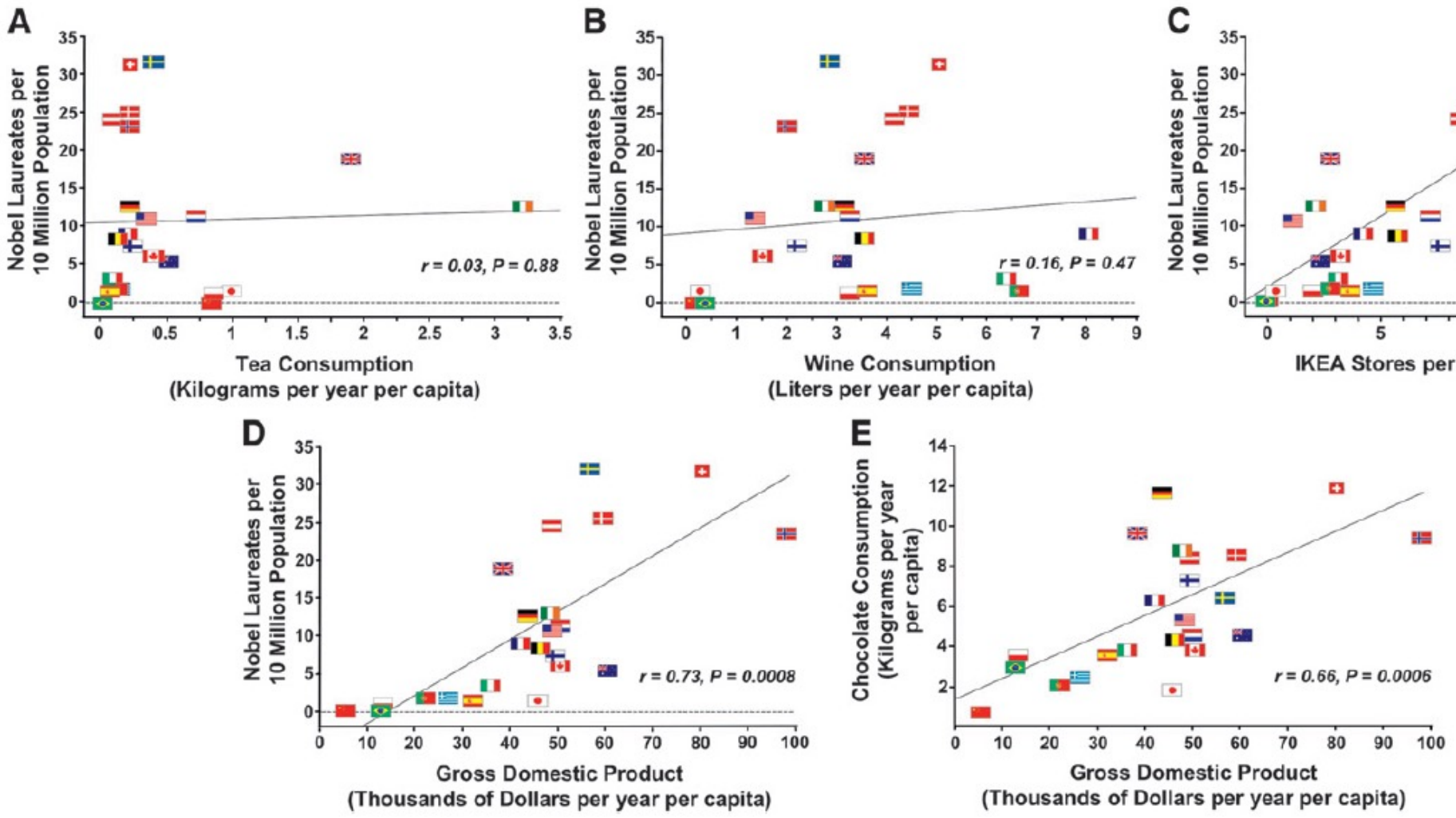
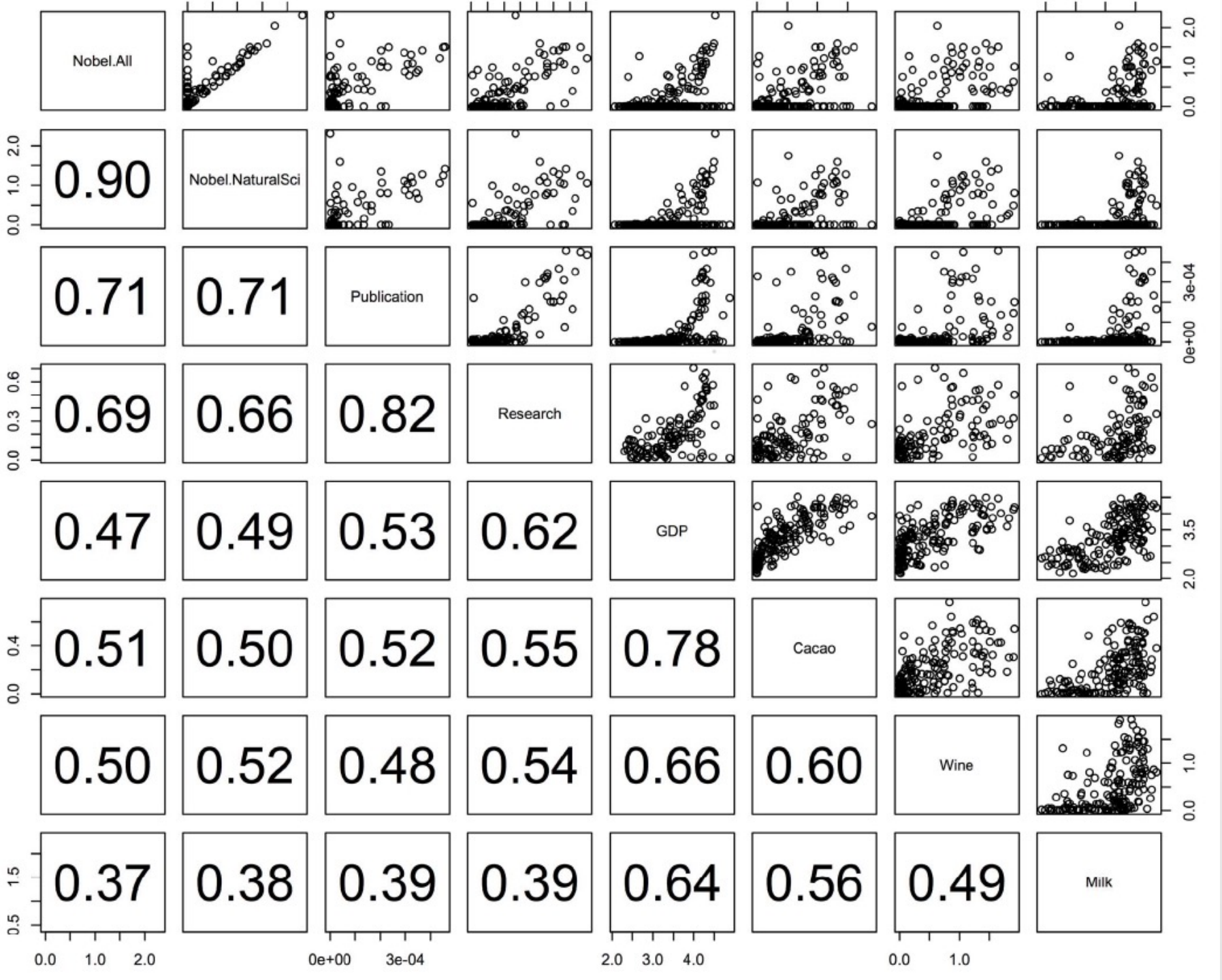


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.


A
Nobel Laureates per 10 Million Population
r = 0.03, P = 0.88
Tea Consumption (Kilograms per year per capita)
B
Nobel Laureates per 10 Million Population
r = 0.16, P = 0.47
Wine Consumption (Liters per year per capita)
C
Nobel Laureates per 10 Million Population
r = 0.82, P < 0.0001
IKEA Stores per 10 Million Population
D
Nobel Laureates per 10 Million Population
r = 0.73, P = 0.0008
Gross Domestic Product (Thousands of Dollars per year per capita)
E
Chocolate Consumption (Kilograms per year per capita)
r = 0.66, P = 0.0006
Gross Domestic Product (Thousands of Dollars per year per capita)


Nobel.All
Nobel.NaturalSci
Publication
Research
GDP
Cacao
Wine
Milk
0.90
0.71 0.71
0.69 0.66 0.82
0.47 0.49 0.53 0.62
0.51 0.50 0.52 0.55 0.78
0.50 0.52 0.48 0.54 0.66 0.60
0.37 0.38 0.39 0.39 0.64 0.56 0.49
0.0 1.0 2.0
0e+00 3e-04
2.0 3.0 4.0
0.0 1.0
0.0 0.3 0.6
0.0 0.4
0.5 1.5
2.0 3.5

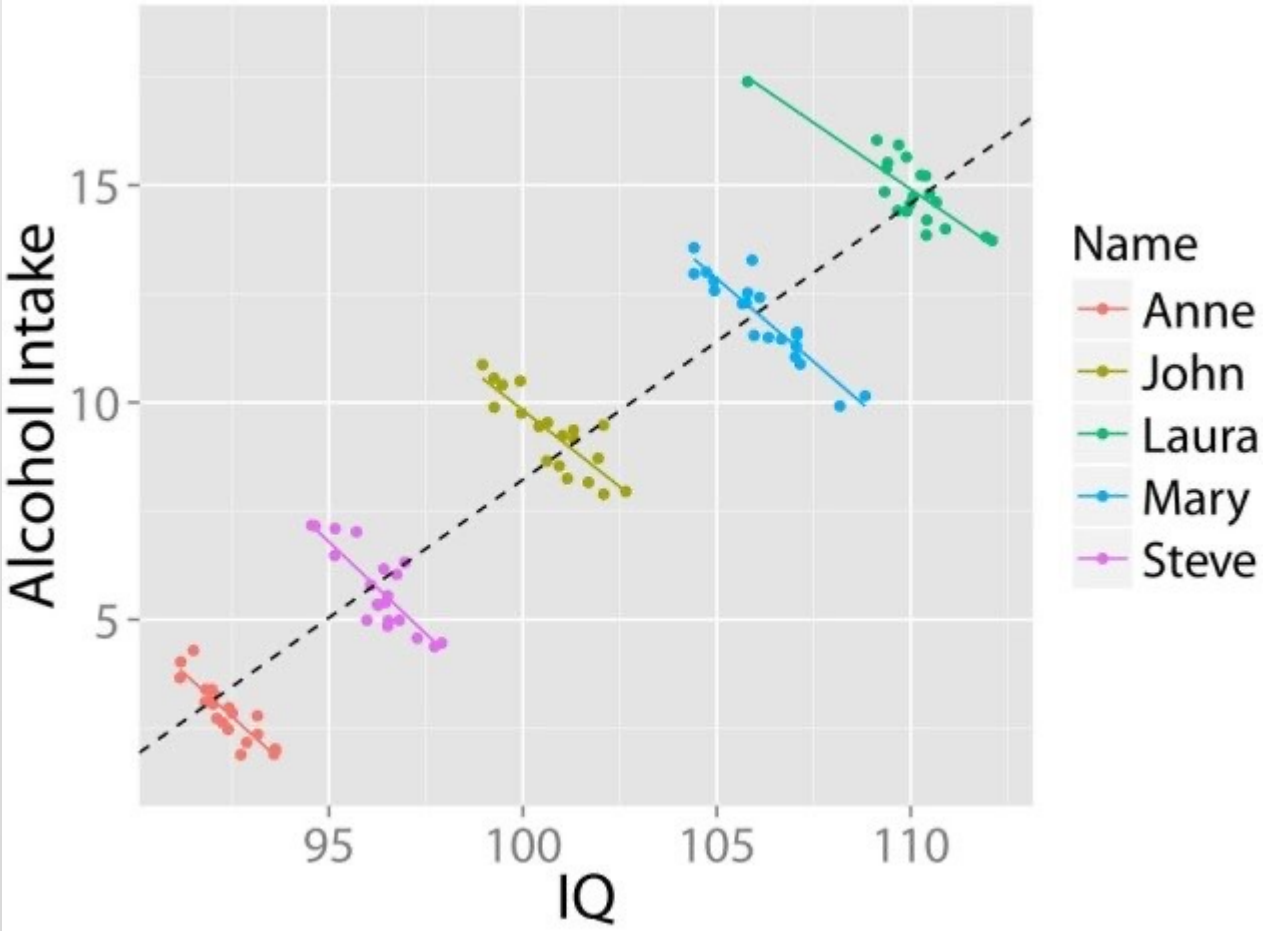
Pe scurt: corelație și erori

- Corelația este un posibil simptom al cauzalității
  - Corelația este observabilă în norul de puncte
  - Coeficienții de asociere măsoară intensitatea relației
- Modelul face legătura dintre corelație și explicație
- Erori de modelare
  - Sensul cauzării
  - A fi corect din rațiuni false
  - Corelații coincidentale
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erorile pot proveni nu doar din model, ci și din eșantionare

Bibliografie

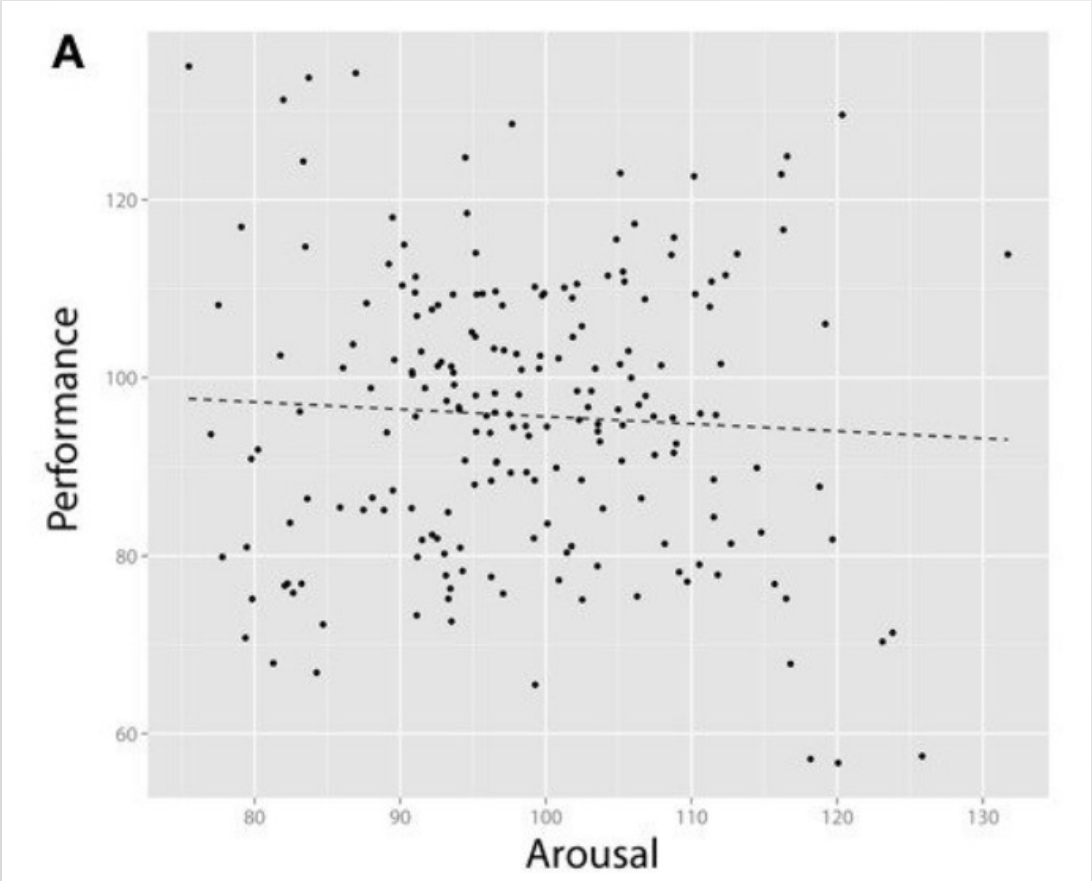
- [1] Vigen, T. (2015). *Spurious correlations*. Hachette books. http://www.tylervigen.com/spurious-correlations
- [2] Steve Borgatti,Simpson's Paradox. https://sites.google.com/site/ba762researchmethods/materials/handouts/simpsonsparadox
- [3] Kievit, R., Frankenhuis, W. E., Waldorp, L., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in psychology*, *4*, 513.
- [4] F. Perry Wilson, Confounding: What Did the Researcher Miss?, https://www.youtube.com/watch?v=HMTs5AZtjpE
- [5] Messerli, F.H. Chocolate Consumption, Cognitive Function, and Nobel Laureates.NEJM. (2012)
- [6] Maurage, P., Heeren, A., & Pesenti, M. (2013). Does chocolate consumption really boost Nobel award chances? The peril of over-interpreting correlations in health studies. *The Journal of Nutrition*, *143*(6), 931-933.
- [7] Doi, H., Heeren, A., & Maurage, P. (2014). Scientific activity is a better predictor of Nobel award chances than dietary habits and economic factors. *PLoS One*, *9*(3), e92612.
- [8] Shona McCombes, An introduction to sampling methods, https://www.scribbr.com/methodology/sampling-methods/
- [9]. Wikipedia contributors. Correlation and dependence. https://en.wikipedia.org/wiki/Correlation_and_dependence
- [10] Zeiger, J. S., Silvers, W. S., Fleegler, E. M., & Zeiger, R. S. (2020). Attitudes about cannabis mediate the relationship between cannabis knowledge and use in active adult athletes. *Journal of Cannabis Research*, *2*, 1-13.

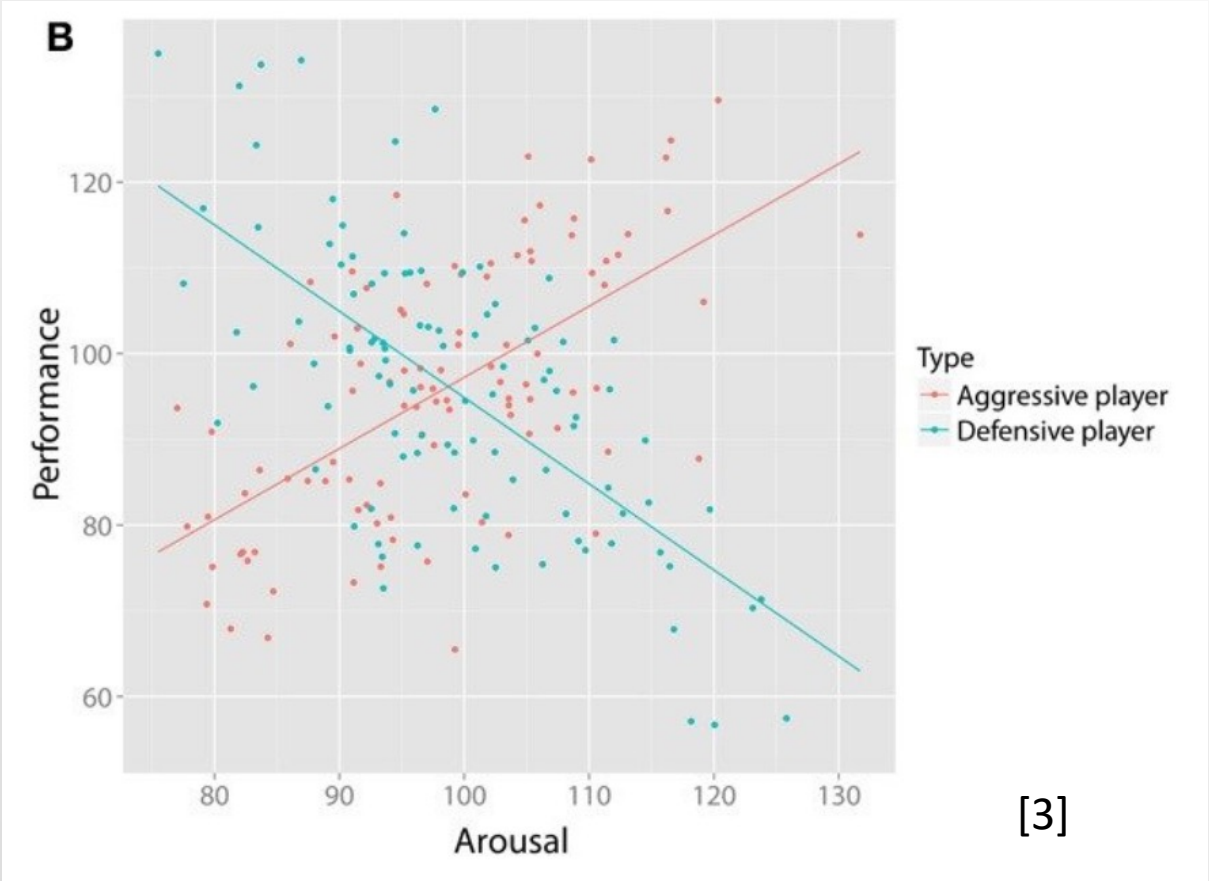
Paradoxul lui Simpson



[3]

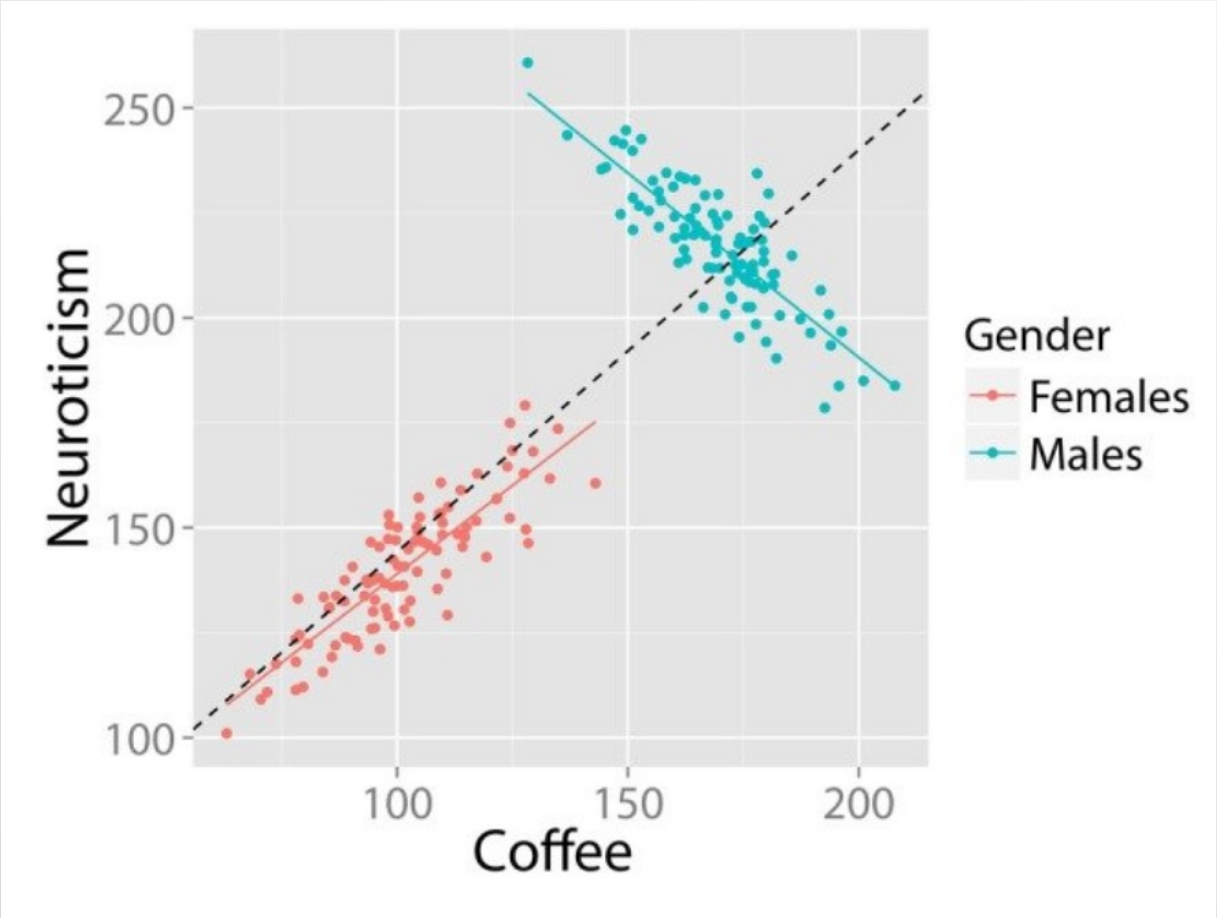
Paradoxul lui Simpson





[3]

Paradoxul lui Simpson



[3]

SYSTEMS LABORATORY

Cursul DS03

Tehnologiile datelor: Programare vs. Învățare automată.
Procesul învățării automate

PR network crunch

Structura cursului

- Introducere: De ce DS?
- Scopul cursului
- Măsurarea și colectarea datelor
- Erori de măsurare
  - Precizia
  - Validitatea
  - Fidelitatea
- Corelația
- Erori de modelare
  - Corelații iluzorii
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erori de eșantionare

- Tehnologiile datelor: programare vs. învățare automată (ML)
- Procesul învățării automate
- Machine Learning și Deep Learning

- Predicție și cunoaștere: analiza datelor
  - Măsurare și reducerea dimensionalității
  - Clasificare (cluster)
  - Cauzalitate și regresie
  - Predicții și serii de timp
- Predicția ca premisă pentru acțiune
- Predicția ca produs economic

DS01: Pe scurt despre știința datelor

- Abilități esențiale privind lucrul cu informații
- Abilități tot mai cerute datorită avansului AI și Big Data
- Două competențe cheie: să ții cont simultan de...
  - Procesul măsurat și procesul măsurării
  - Tendința centrală, variabilitate și excepții
- Tendința centrală e estimată prin: medie, mediană, mod
- Variabilitatea: variația, deviația standard
- Erorile de măsurare pot privi:
  - Precizia: sistematice / aleatoare
  - Validitatea
  - Fidelitatea (repetabilitatea)

DS02: Pe scurt despre corelație și erori

- Corelația este un posibil simptom al cauzalității
  - Corelația este observabilă în norul de puncte
  - Coeficienții de asociere măsoară intensitatea relației
- Modelul face legătura dintre corelație și explicație
- Erori de modelare
  - Sensul cauzării
  - A fi corect din rațiuni false
  - Corelații coincidentale
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erorile pot proveni nu doar din model, ci și din eșantionare

Cuprins

3.1 De ce ML?

3.2 Tipuri de probleme pentru ML

A. Învățarea supravegheată

1. Clasificare
2. Regresie & serii de timp
3. Sequence to Sequence

B. Învățarea nesupravegheată

1. Clustering
2. Reducerea dimensionalității
3. Transfer learning
4. Reinforcement learning

3.3 Procesul învățării automate

3.4 Auto ML

#3.1 De ce ML?

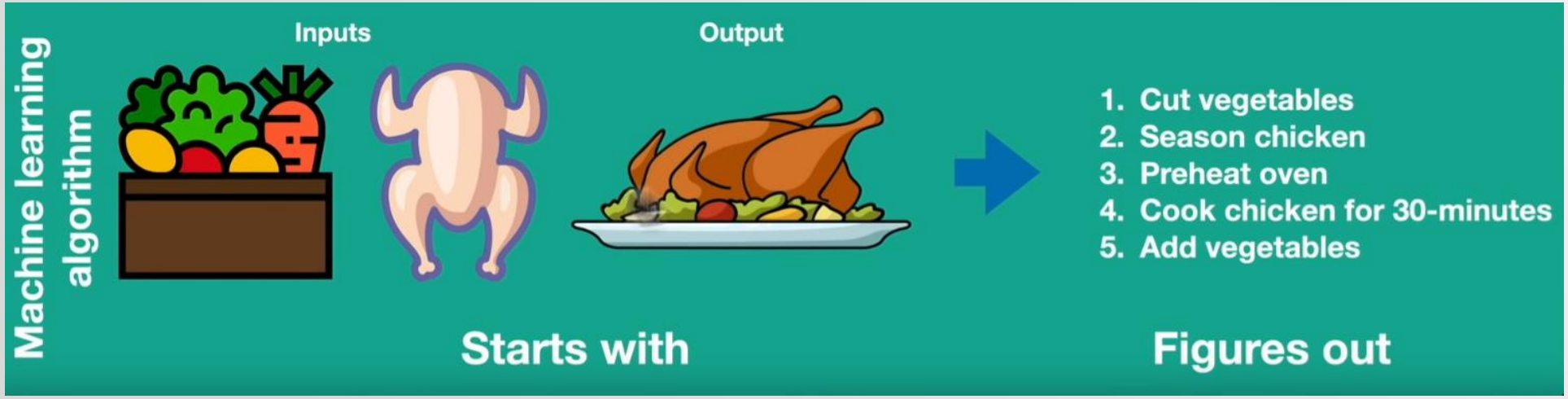
Ce este învățarea automată?

La ce este bună?

Învățarea automată

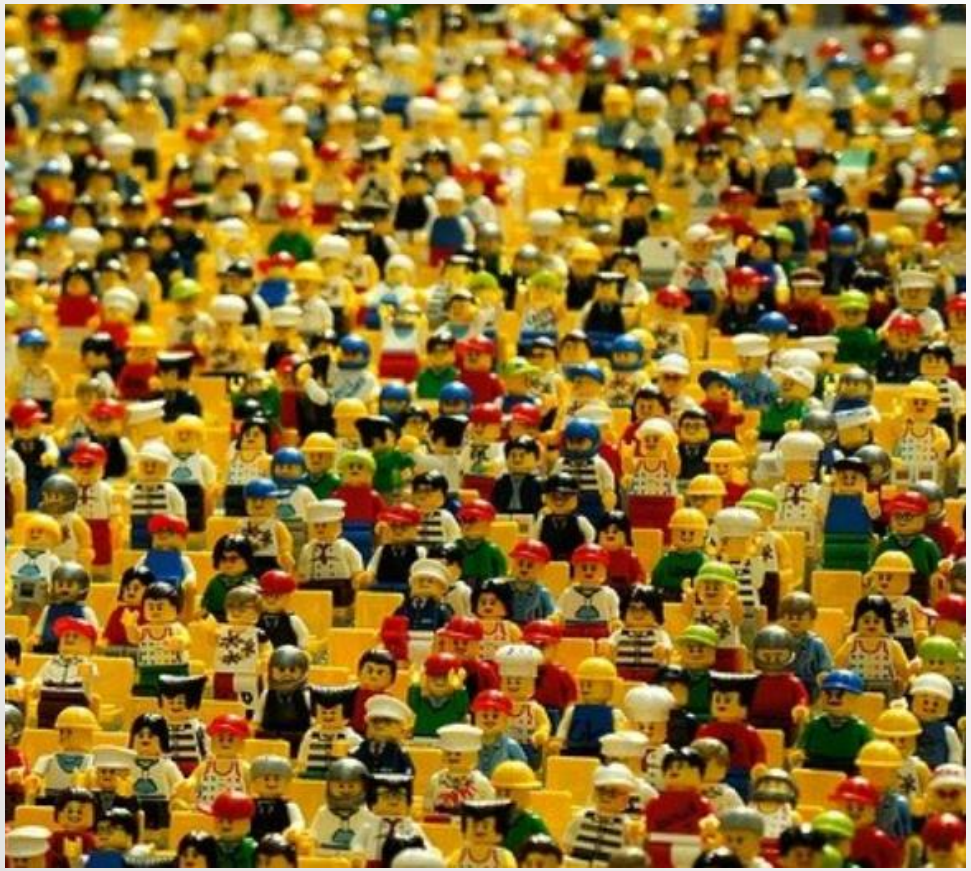
- Procesul de a transforma **date** în **numere**, și de a găsi tipare/structuri în acele numere.
- Învățarea automată oferă calculatoarelor abilitatea de a învăța, fără să fie programate în mod explicit

Programare vs. Învățare automată



De ce învățarea automată?

- Probleme foarte complexe pe volume foarte mari de date
- Nu există reguli clare ce pot fi căutate, găsite și implementate algoritmic
- Exemplu
  - Segmentarea unei populații (Clustering vs. Classification)
  - Mașini autonome



Reguli prezente vs. posibile

- Dacă un sistem bazat pe reguli/euristici poate rezolva problema, nu folosiți învățare automată! ([Rule #1 of Google's handbook for ML](#))

ML – când?

- **Probleme unde lista de reguli pare că nu se mai termină**
  - ML vine și încearcă să rezolve problemele unde programarea tradițională nu reușește.
- **Sisteme în continuă schimbare**
  - Față de programele clasice, ML se poate adapta unor scenarii noi și poate evolua 'aproape singură' cu trecerea timpului.
- **Înțelegerea unor volume imense de date**
  - Clasificarea a milioane de articole / recenzii necategorizate necesită o forță de muncă umană copleșitoare.

#3.2 Tipuri de probleme

A. Învățare supravegheată

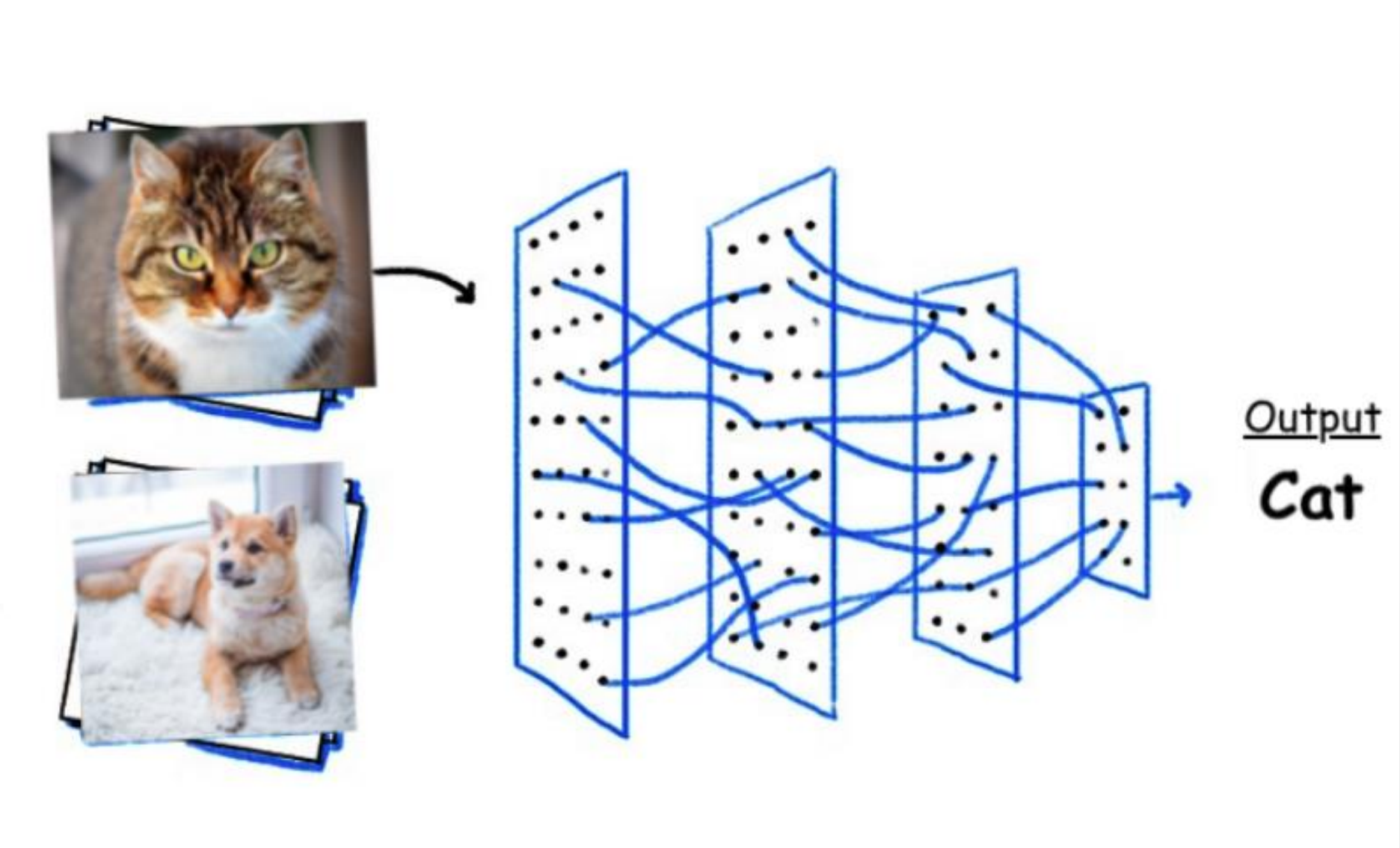
B. Învățare nesupravegheată

A. Învățare supravegheată

- Îi oferim modelului exemple, și îi spunem la fiecare pas cât de bine a învățat
- Exemple:
  1. Clasificare
  2. Regresie & serii de timp
  3. Sequence to Sequence

A.1. Clasificare

- Transformăm o **secvență** (text, imagini, sunet, etc) în **indici** de categorie.
- Vrem să asociem fiecărei intrări una sau mai multe categorii finale:
  - Comentariul este pozitiv sau negativ? (clasificare binară)
  - Ce rasă de câine? (multi-class)
  - Care sunt genurile unei film? (multi-label)


Output
Cat

SYSTEMS LABORATORY

MIT Technology Review

Artificial intelligence Dec 20

A US government study confirms most face recognition systems are racist

Almost 200 face recognition algorithms—a majority in the industry—had worse performance on nonwhite faces, according to a landmark study.

PR network crunch

MIT Technology Review

Regresie

- Prezicem valori numerice continue (necunoscute) pe baza corelației cu alte variabile cunoscute
  - Cât costă un apartament cu 3 camere din centru (preț vs. zonă)?
  - Cu cât crește consumul de carne în funcție de GDP/capita?

Meat consumption vs. GDP per capita, 2017

Our World in Data

Average meat consumption per capita, measured in kilograms per year versus gross domestic product (GDP) per capita measured in 2011 international-$. International-$ corrects for price differences across countries. Figures do not include fish or seafood.

Source: UN FAO; World Bank, World Development Indicators

OurWorldInData.org/meat-production • CC BY

PR network crunch

Sursă

Serii de timp

- Prezicem valori numerice continue (necunoscute) pe baza evoluției lor până acum
  - Câte grade vor fi mâine?
  - Care va fi prețul unui Bitcoin peste două zile?
- Combină
  - Proprietăți temporale interne: cicluri, tendințe
  - Modele explicative: cauze externe

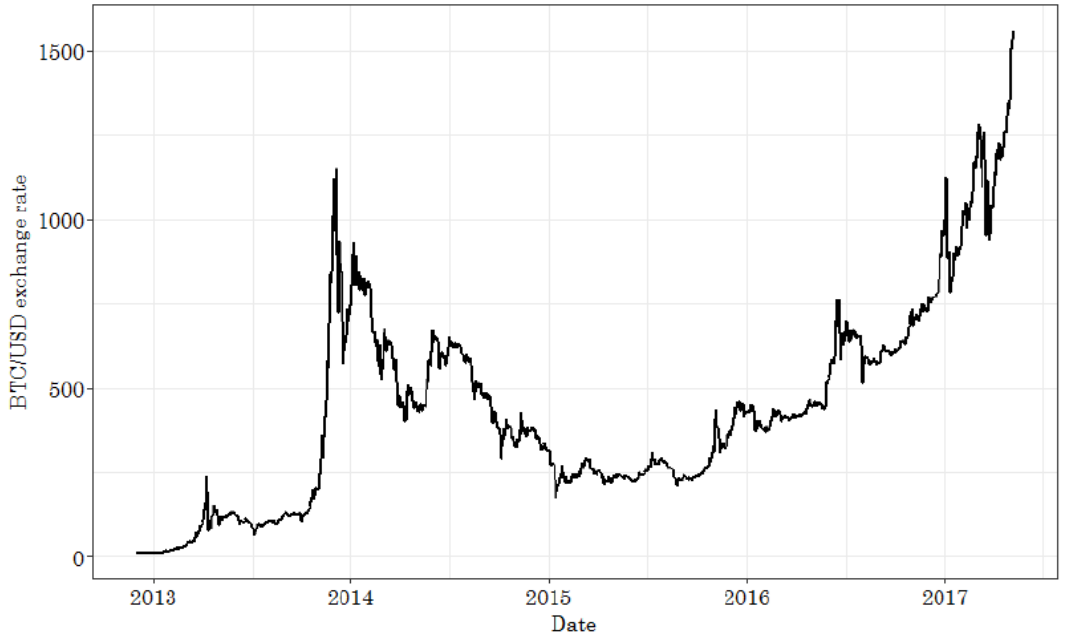
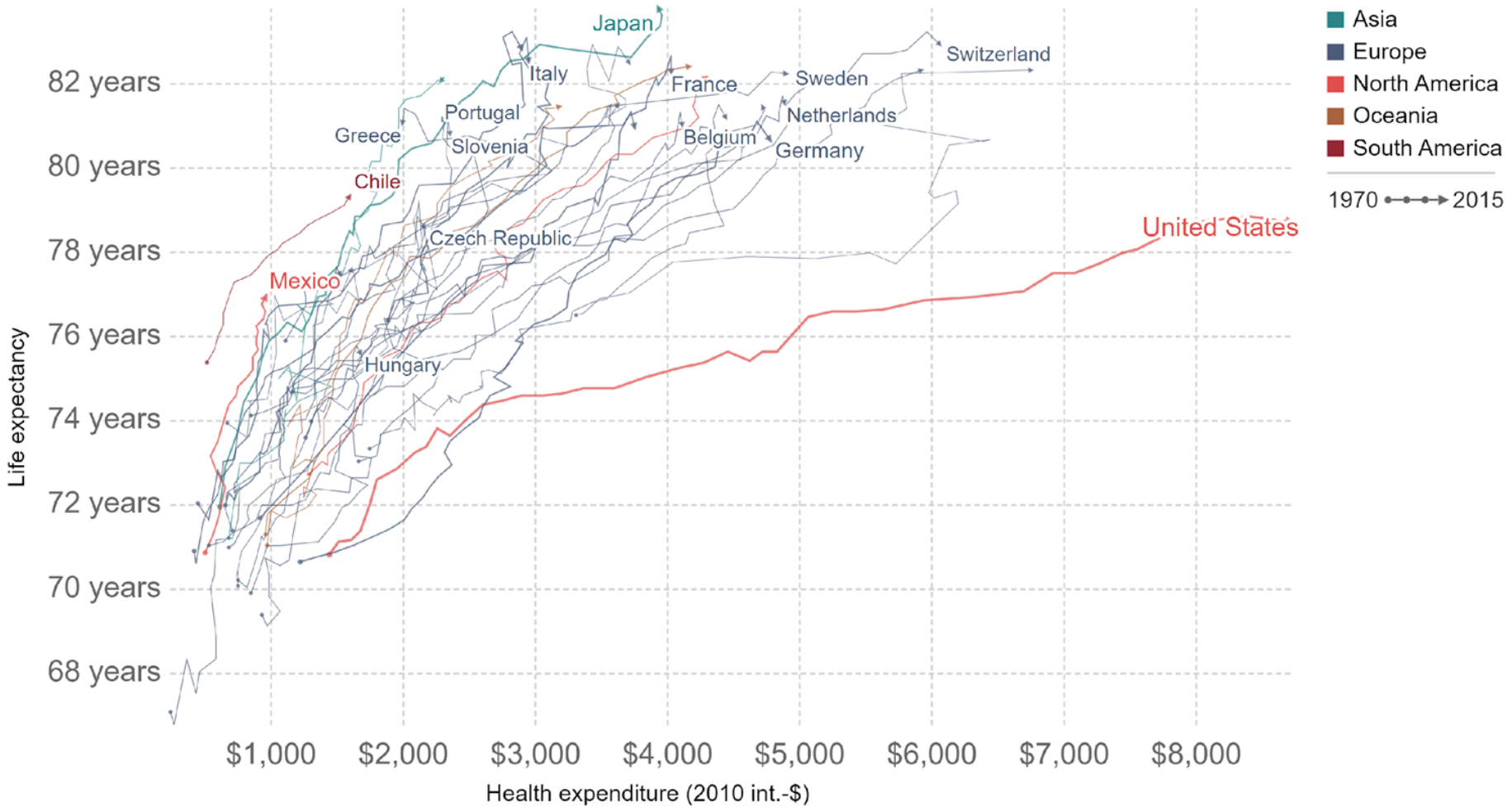


Figure 2: Bitcoin exchange rate with USD

Poyser 2018

Life expectancy vs. health expenditure, 1970 to 2015

Health financing is reported as the annual per capita health expenditure and is adjusted for inflation and price level differences between countries (measured in 2010 international dollars).



Source: World Bank, Health Expenditure and Financing - OECDstat (2017), Population (Gapminder, HYDE(2016) & UN (2019)), Our World In Data

OurWorldInData.org/the-link-between-life-expectancy-and-health-spending-us-focus • CC BY

PR network crunch

Sursă

Sequence to sequence

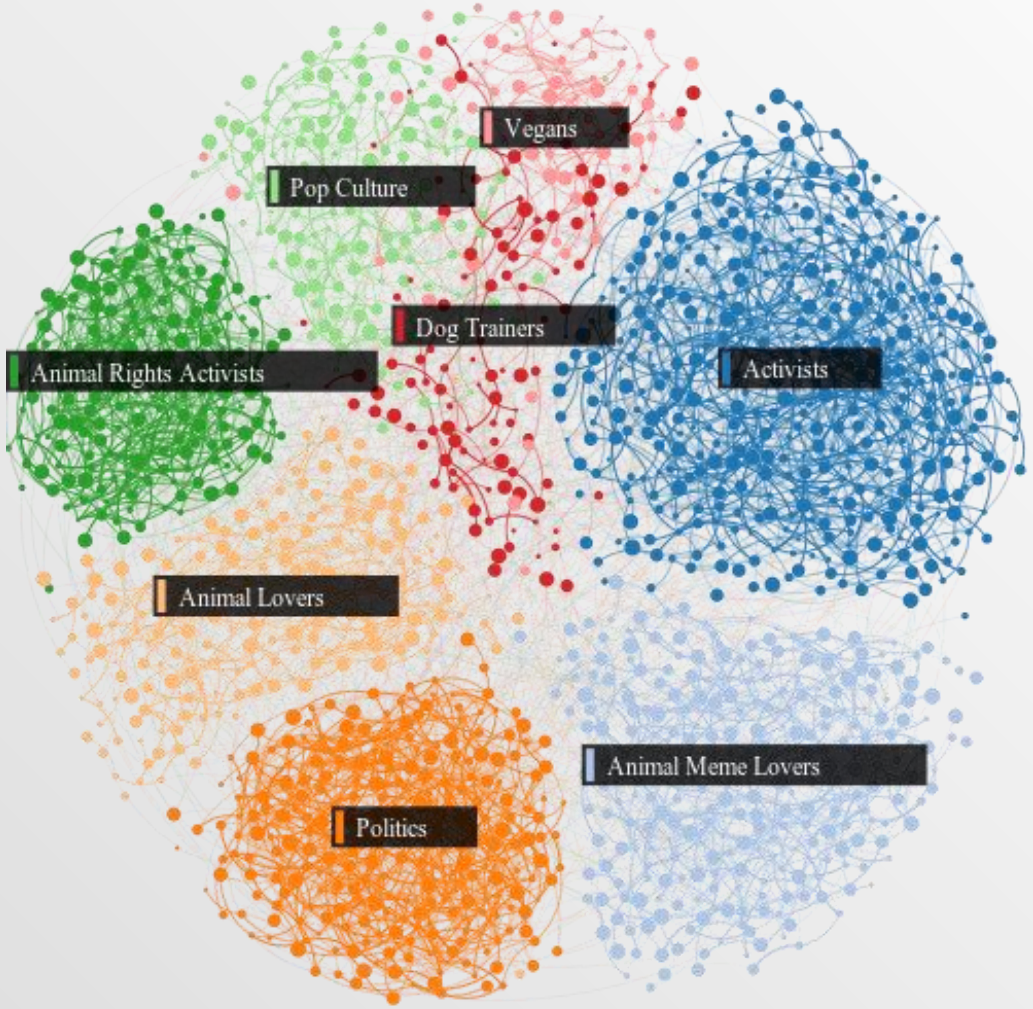
- Primim o secvență de date, și întoarcem tot o secvență de date.
  - Traducerea unei propoziții într-o altă limbă.
  - Transformarea vocii în text.
  - Răspunderea la o întrebare anume.
  - Rezumarea unui articol lung.
  - Extragerea unei propoziții doar din imagini cu mișcarea buzelor unei persoane.

B. Învățare nesupravegheată

- Nu avem date organizate, și încercăm să lăsăm modelul să creeze o structură
- Exemple:
  1. Clustering
  2. Reducerea dimensionalității
  3. Transfer learning
  4. Reinforcement learning

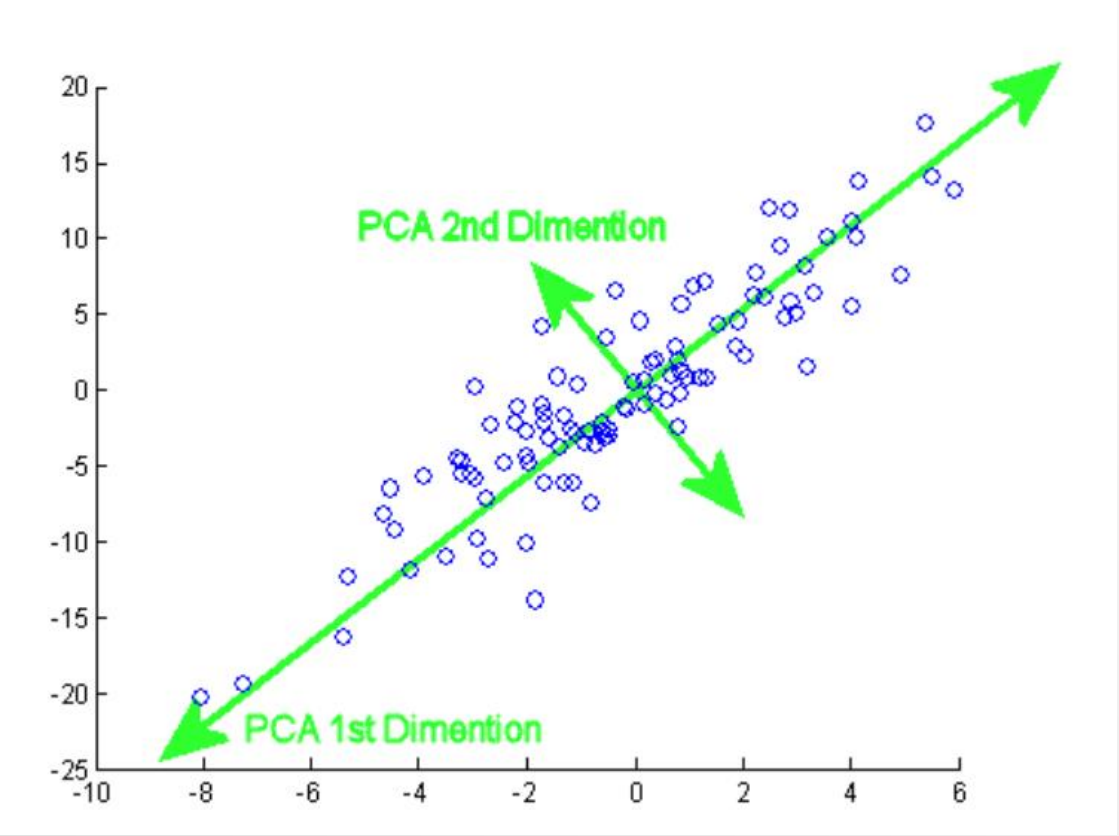
B.1 Clustering

- Crearea unei reprezentări vectoriale a tuturor datelor și încercarea de a găsi grupuri similare în acestea.
  - Cum organizez clienții unei linii aeriene în funcție de interesele lor exprimate pe Twitter? (Davies 2019)



B.2 Reducerea dimensionalității

- Care sunt cele mai importate dimensiuni ale datelor mele?
  - Compresia datelor
  - Descoperirea structurilor fundamentale
  - Eliminarea erorilor de măsurare
  - Vizualizarea volumelor uriașe de date
  - Extragerea de date pentru antrenarea altor modele



Sursă

B.3 Transfer learning

- Antrenarea (îndelungată) unui model pe un task general, și folosirea ulterioară a lucrurilor învățate pentru probleme similare.
  - Transformers
  - Style-transfer models

Transformers

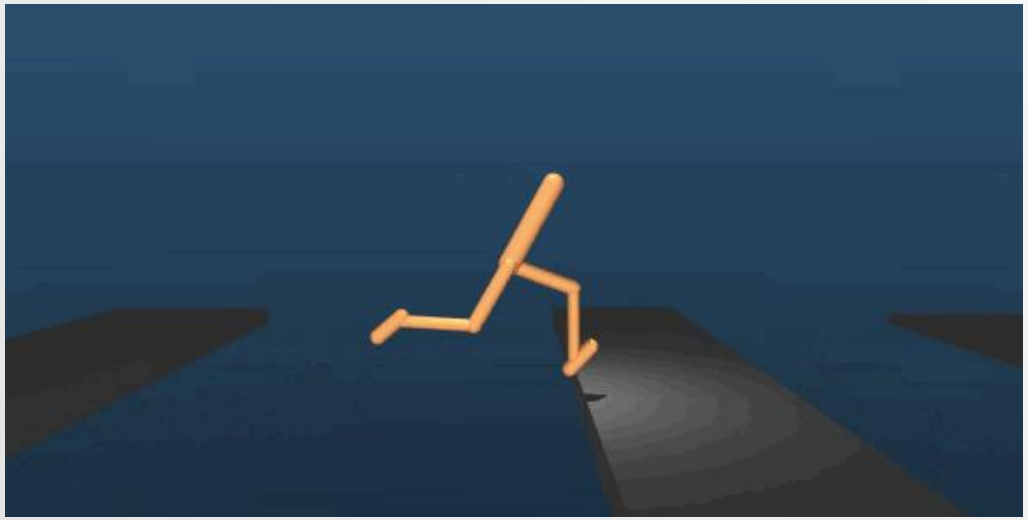
- Modele antrenate pe corpuri imense de text (întreaga Wikipedia + zeci de mi de cărți), care sunt apoi folosite pentru antrenarea rapidă problemelor de procesare a limbajului natural. (Giancaglia 2019)

Style transfer models

- Înțelegerea *stilului* unei imagini, și aplicarea lui asupra unei imagini noi

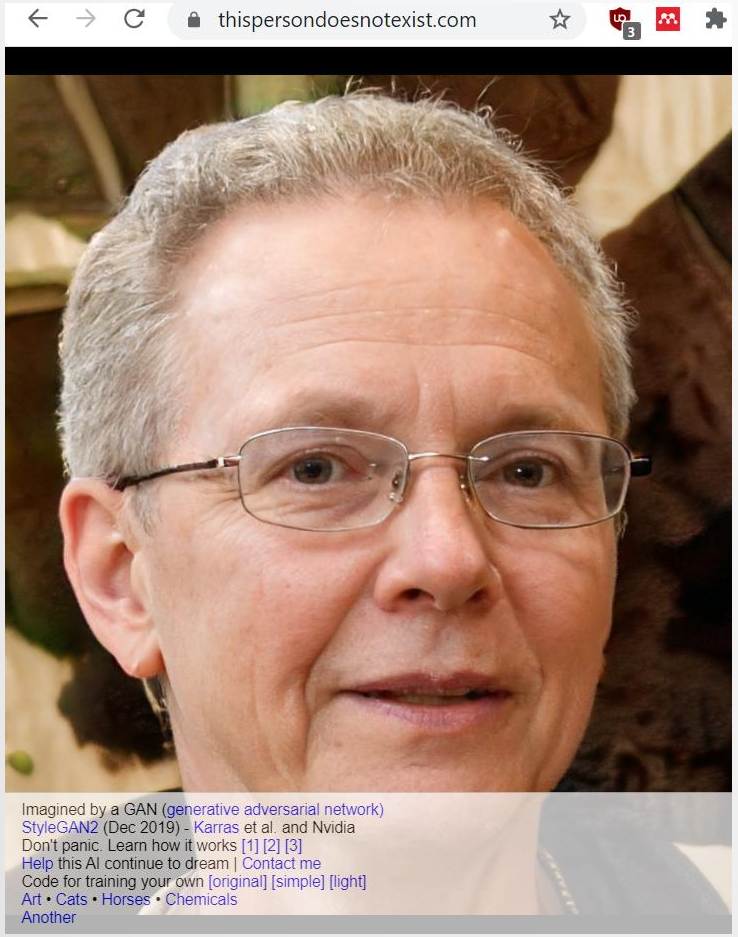
B.4 Reinforcement learning

- Poate rezolva probleme în care nu știm pas cu pas dacă o acțiune este corectă, și avem ca scop maximizarea unui rezultat final.
- Învățarea unui agent să meargă.
  - Știm să mergem, dar putem descrie pas cu pas cum o facem? Recompensăm modelul (doar la **final**!) pentru fiecare metru pe care îl parcurge și îl lăsăm să învețe singur din propriile greșeli!



Modele adversariale

- Putem avea și modele **adversariale** (generator și discriminator) care învață simultan!
  - Model de identificare a fețelor umane și model de generare a fețelor umane. Se poate crea un ciclu închis de învățare, și ambele devin mai bune.
  - Model care ascunde cuvinte cheie dintr-o propoziție și model care ghicește ce cuvinte ar trebui să fost acolo.



thispersondoesnotexist.com

Sursă

SYSTEMS LABORATORY

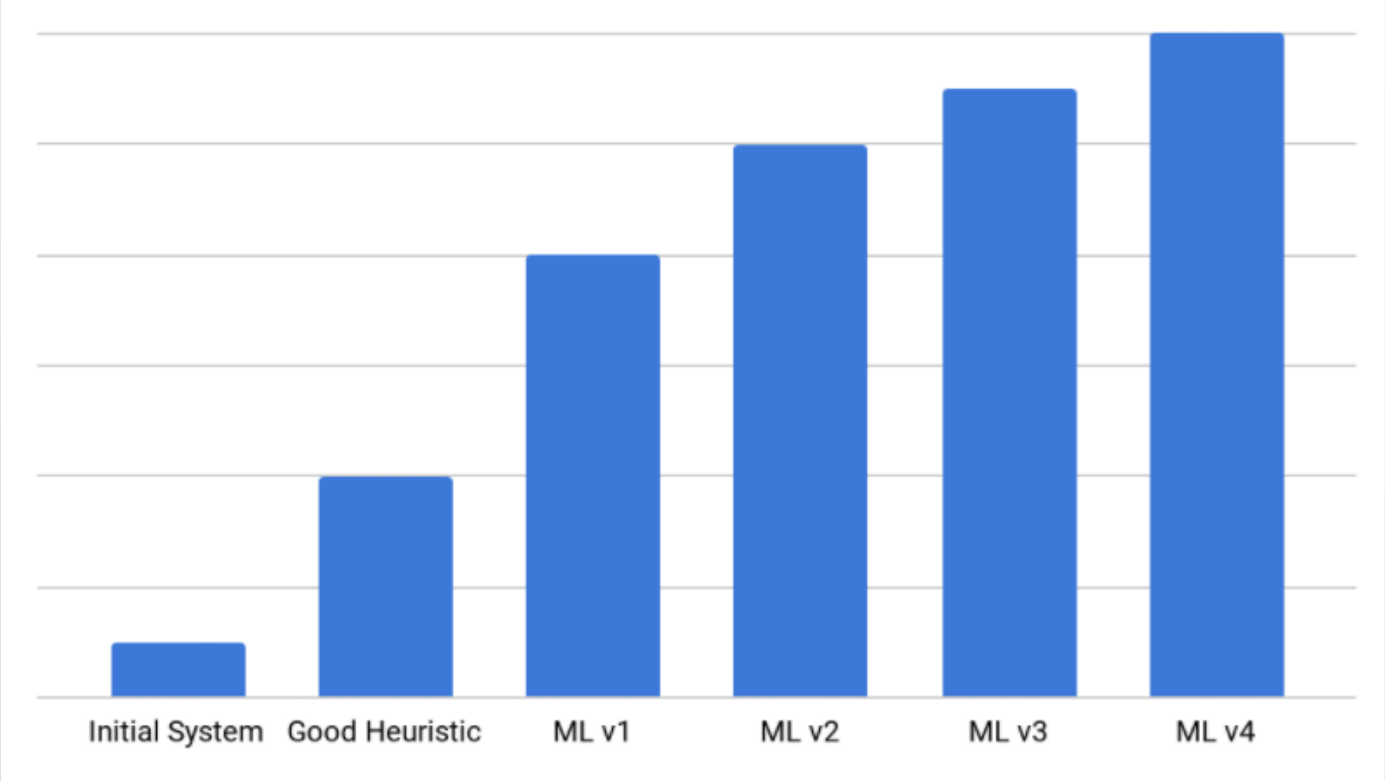
#3.3 Procesul de învățare automată

De la formularea problemei la evaluarea modelului, punerea în producție și re-antrenarea lui

PR network crunch

Sigur am nevoie de învățare automată?

- Mereu, dacă este posibil, începe cu o euristică care rezolvă problema.
- Apoi, iterează diferite modele de ML care încearcă să o bată.

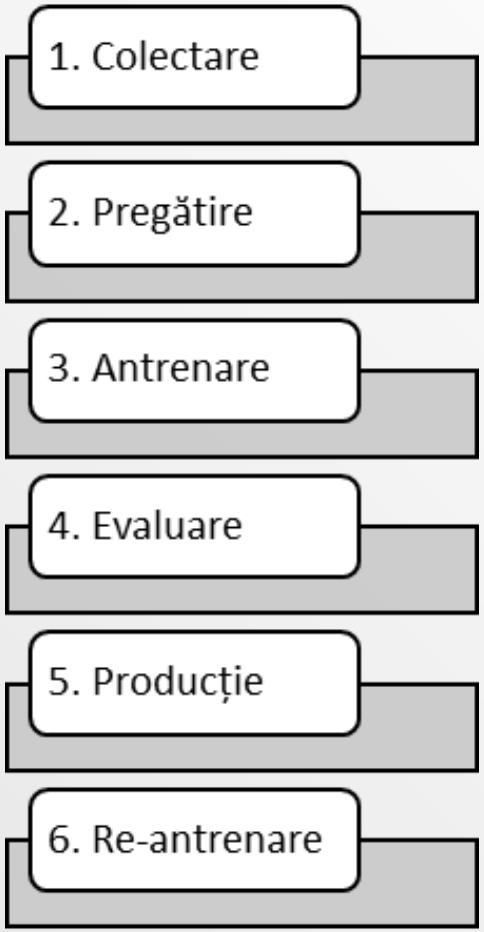


Etape ale formulării problemei

1. **Scop**: Ce vreau să facă modelul meu?
   - *Câte vânzări voi face astăzi? Simplă euristică: Fac o medie a vânzărilor din ultimii ani în această zi.*
2. **Obiectiv**: Care este rezultatul ideal?
   - Cum știu că modelul meu are o performanță bună?
   - Care sunt metricile?
   - Sunt metricile adecvate pentru performanța reală a modelului?
   - Acoperă metricile toate cazurile?
3. **Date**: Unde pot găsi date? Există deja? Pot genera?
   - Există anumite părtiniri (*bias*) în datele mele?

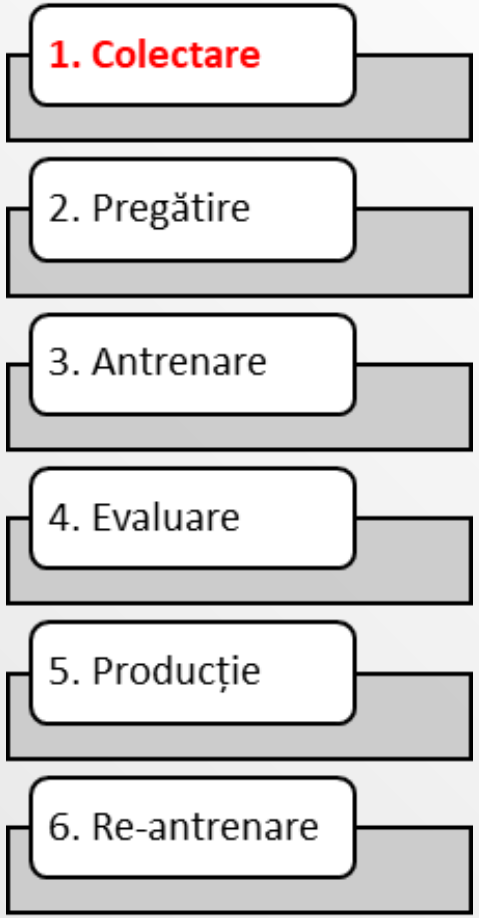
Procesul de învățare automată

1. Colectarea datelor
2. Pregătirea datelor
3. Antrenarea
4. Evaluarea modelului
5. Producția: aplicarea modelului
6. Re-antrenarea modelului



1. Colectarea datelor

- Exact ce **tip** de problemă încerc să rezolv?
- Există seturi/surse de **date deja existente** pentru problema mea?
- Trebuie să îmi pun problema de **privacy**?
- Există o **structură** în date? Sau trebuie să o găsesc?



2.1 Pregătirea datelor

- **Explorarea datelor**
  - Grafice/word clouds/statistici
  - Tendință centrală, variabilitate, excepții
- Cum pot reprezenta **numeric** datele?
  - Le pot structura mai bine? (data engineering)
- **Curse of dimensionality**: am prea multe date?
  - Cum păstrez doar esențialul?
  - Reducerea dimensionalității

1. Colectare
2. Pregătire
3. Antrenare
4. Evaluare
5. Producție
6. Re-antrenare

2.2 Pregătirea datelor

- Adaugă reprezentarea mea numerică un **bias**?
  - Pixelii luminați au valori mai mari ca cei închiși: pot 'valora' mai mult
  - Dacă reprezint țări prin numere (Spania - 1, România - 2) înseamnă că România > Spania? (soluție: *one-hot encoder*)

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

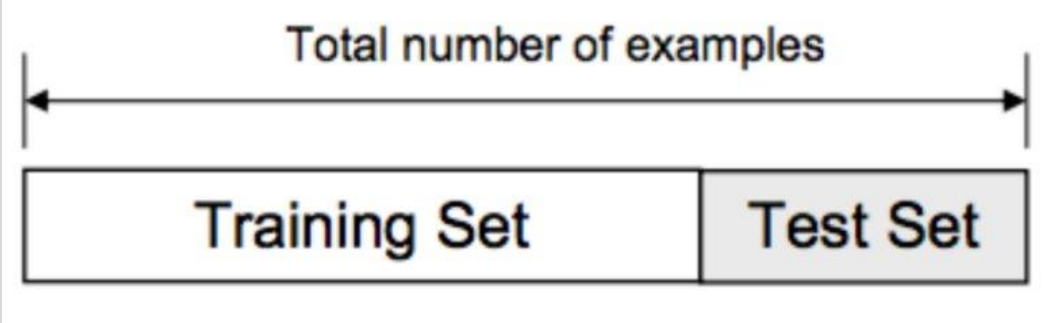
One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

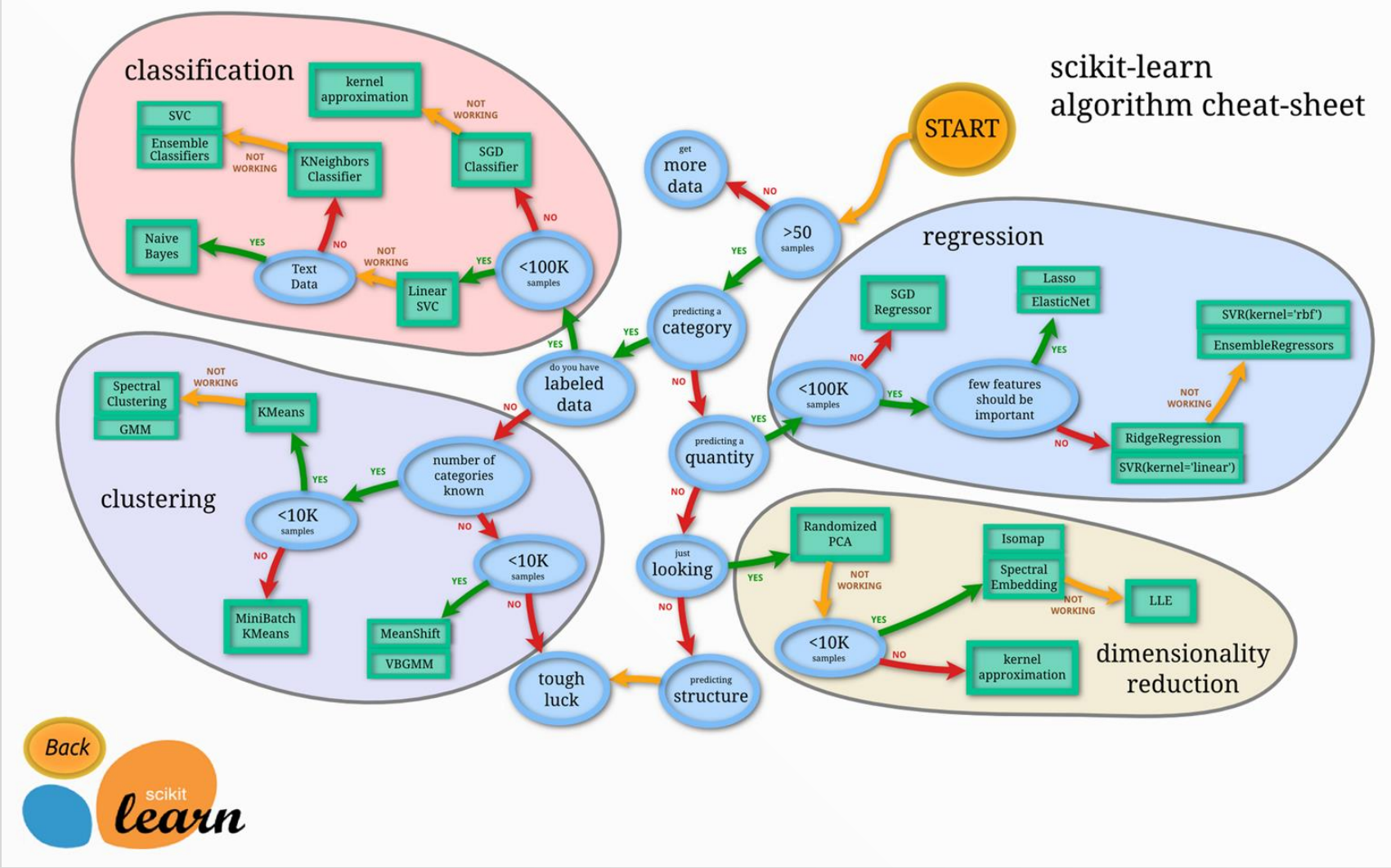
1. Colectare
2. **Pregătire**
3. Antrenare
4. Evaluare
5. Producție
6. Re-antrenare

2.3 Pregătirea datelor

- Împărțirea datelor
  - Dacă îmi folosesc toate datele pentru antrenare, pe ce mai fac testarea?
  - Ex.: Îmi împart datele în **80%** pentru antrenare, **20%** pentru test.



3.1 Antrenarea: alegerea modelului



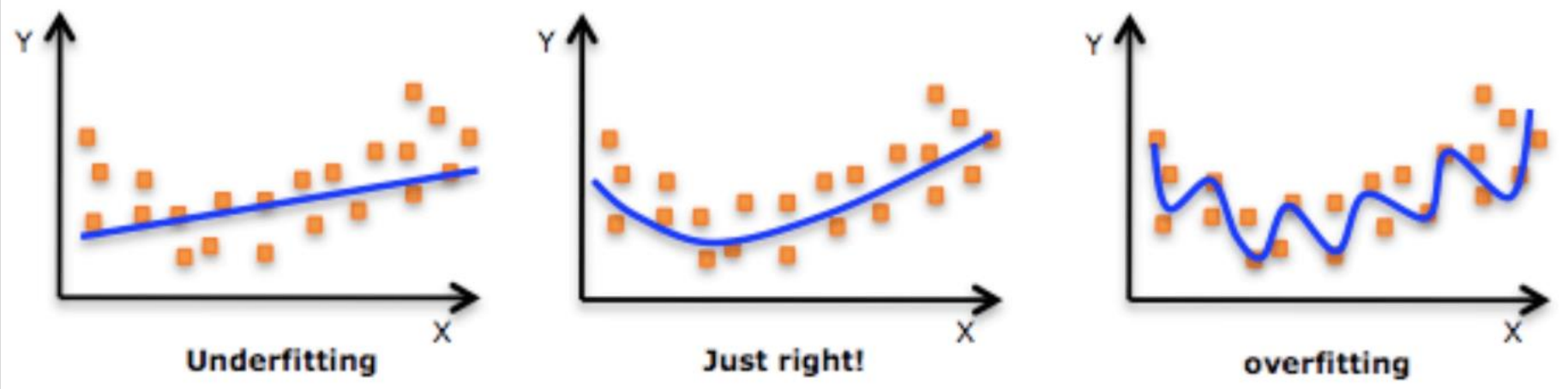
1. Colectare
2. Pregătire
3. **Antrenare**
4. Evaluare
5. Producție
6. Re-antrenare

3.2 Antrenarea: tipuri

- **Batch learning**: Am deja un set static și mare de date, îl antrenez și îl lansez în producție.
- **Continuous learning**: Există un flux de date, iar modelul este actualizat în mod continu. *Antrenările sunt mici și dese. Învață chiar și după ce a fost lansat în producție.*
- **Fine-tuning**: Pornesc de la un model deja antrenat pe un caz general asemănător problemei mele, și îl antrenez pe cazul meu specific. *Este o modalitate foarte rapidă și necostisitoare de a avea rezultate SOTA.*

3.3 Antrenarea: provocări

1. Underfitting
2. Overfitting
3. Hiperparametri
4. Viabilitatea în lumea reală



1. Colectare
2. Pregătire
3. **Antrenare**
4. Evaluare
5. Producție
6. Re-antrenare

3.3.1 Underfitting

- Modelul nu are rezultate bune nici pe setul de antrenament
- Modelul nu poate învăța problema
  - Încercăm să antrenăm pentru o perioadă mai lungă de timp.
  - Încercăm să oferim date mai curate și mai utile modelului
    - *GIGO: garbage-in → garbage-out*
  - Adăugăm complexitate modelului, îl facem mai mare, pentru a putea 'înțelege' problema mai bine

3.3.2 Overfitting

- Modelul are rezultate foarte bune pe setul de antrenament, dar la evaluare este considerabil mai prost.
  - Știe perfect toate slide-urile de curs, dar a luat notă mică la examen 😂
- În acest caz modelul a învățat să **replice** datele, și nu **generalizează**.

3.3.2 Overfitting: regularizări

- **Early stopping**: Ne oprim din antrenare când modelul pare că se descurcă din ce în ce mai **prost** pe **setul de testare**, și din ce în ce mai **bine** pe **setul de antrenare**.
- **Batch-normalization:** Standardizăm (re-centrăm și re-scalăm) intrările unui layer pentru fiecare batch.
- **Dropout**: Închidem unele 'căi' ale modelului, cu care el e obișnuit, pentru a îl forța să găsească unelei noi.
  - Vrei ca bunicul să ia și alt drum spre magazin? Blochează drumul vechi!
- **L1 & L2 regularization**: Moduri matematice de a descuraja învățarea unei complexități foarte ridicate.
- **Data augmentation**: Crearea de date artificiale cu scopul de a ne generaliza modelul.
  - *Pentru imagini, le putem roti, oglindi, schimba contrastul..,*
  - *Pentru text putem înlocui cuvinte cu sinonime, reordona, traduce în altă limbă și înapoi*
  - *Pentru sunet, putem schimba pitch-ul, adăuga sunet de fundal, ecou*

3.3.3 Hiperparametri

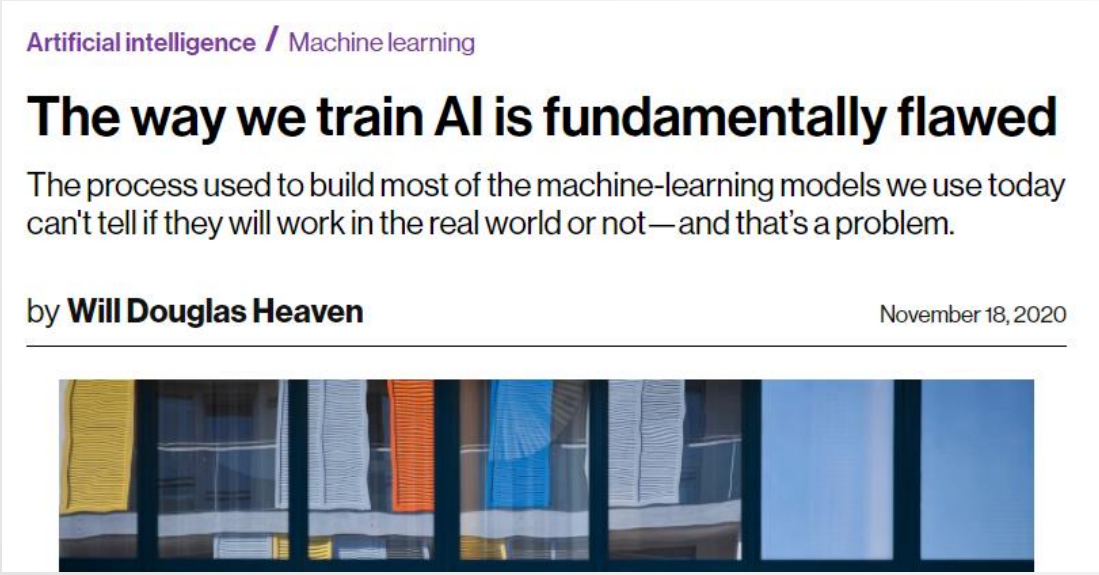
- Structura modelelor noastre este determinată de un număr mare de variabile,
  - Valoarea lor optimă nu este deloc evidentă
- Valoarea parametrilor însă are un impact foarte mare asupra performanței modelului
  - Atât în procesul de antrenare, cât și în producție

3.3.3 Hiperparametri: regularizări

- Modificăm **forma** input-ului și output-ului pentru a le adapta problemei noastre.
- Testăm cazuri **foarte simple** și vedem dacă le poate învăța. Apoi continuăm să adăugăm complexitate.

3.4 Viabilitatea în lumea reală

- Modelele de ML dezvoltate în laborator pot sa nu funcționeze în lumea reală
  - Data shift: Datele de antrenament nu sunt reprezentative
    - Eg.: radiografii de antrenament mult mai clare decât în clinici
  - Sub-specificare: multe modele cu diferențe aparent mici trec testul
    - Aceste mici diferențe devin mari în lumea reală



Artificial intelligence / Machine learning

The way we train AI is fundamentally flawed

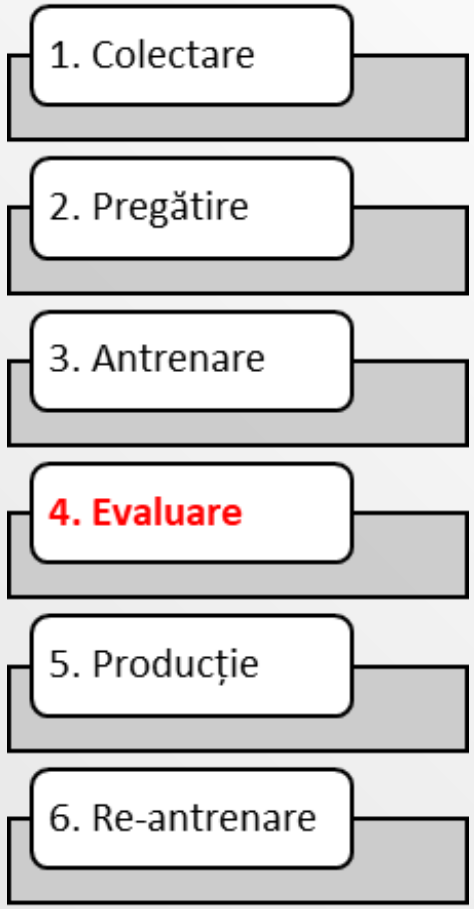
The process used to build most of the machine-learning models we use today can't tell if they will work in the real world or not—and that's a problem.

by **Will Douglas Heaven** November 18, 2020

Sursă

4. Evaluarea modelului

- Folosesc setul de **test** și **metricile** alese pentru a-mi evalua modelul.
  - Fără set de testare, nu pot ști cât de bine se descurcă modelul meu pe exemple noi
  - Dacă metrica de testare este înșelătoare, pot avea impresia că modelul meu este mai bun sau mai prost decât în realitate!
  - Identific ieșirile cu **performanță proastă** pentru a le analiza

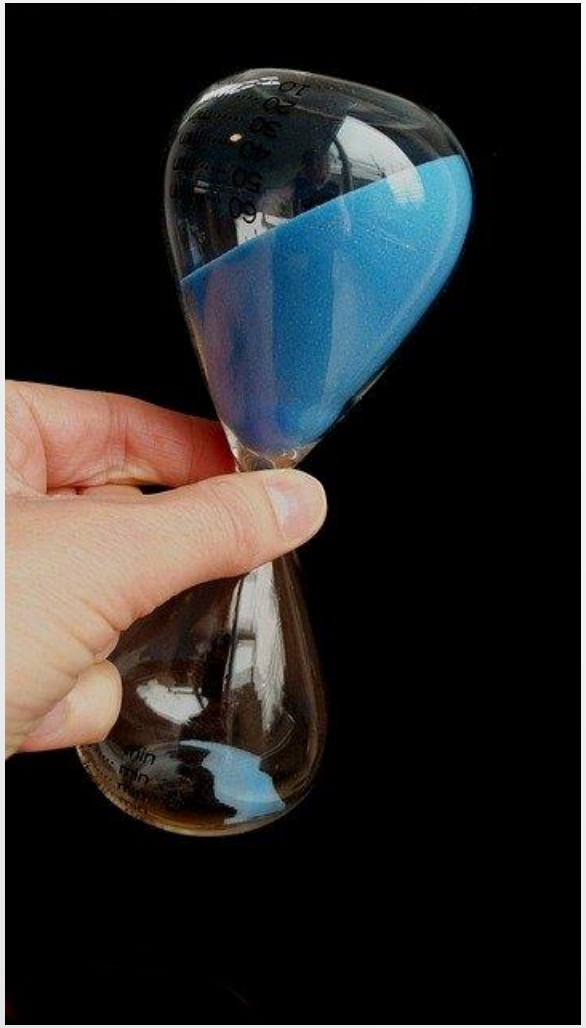


4.1 Evaluarea modelului – cazuri atipice

- Modelul probabil se va descurca bine pe exemplele populare, însă prost pe cele rare / atipice
  - Trebuie să acord o grijă specială cazurilor rare
- Identific cele mai **importante** intrări
  - Pot să mai elimin din restul

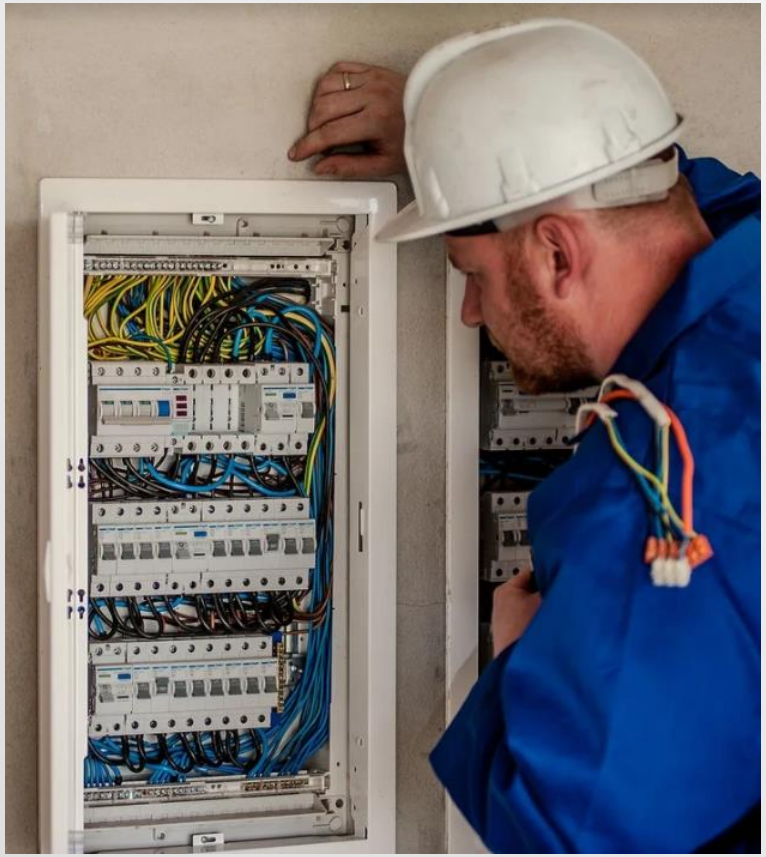
4.2 Evaluarea duratelor

- Analizez **timpul** de antrenare și predicție.
  - *Este fezabil să aștept atât de mult la fiecare antrenare? (este foarte greu să iterezi prin versiuni ale modelului, dacă durează ~1 săptămână să se antreneze)*
  - *Este timpul de predicție suficient de scurt?*
- Aleg un compromis între **acuratețe și timp de predicție**
  - Reproiectez modelul pentru a fi mai complex sau mai simplu.



5. Producția

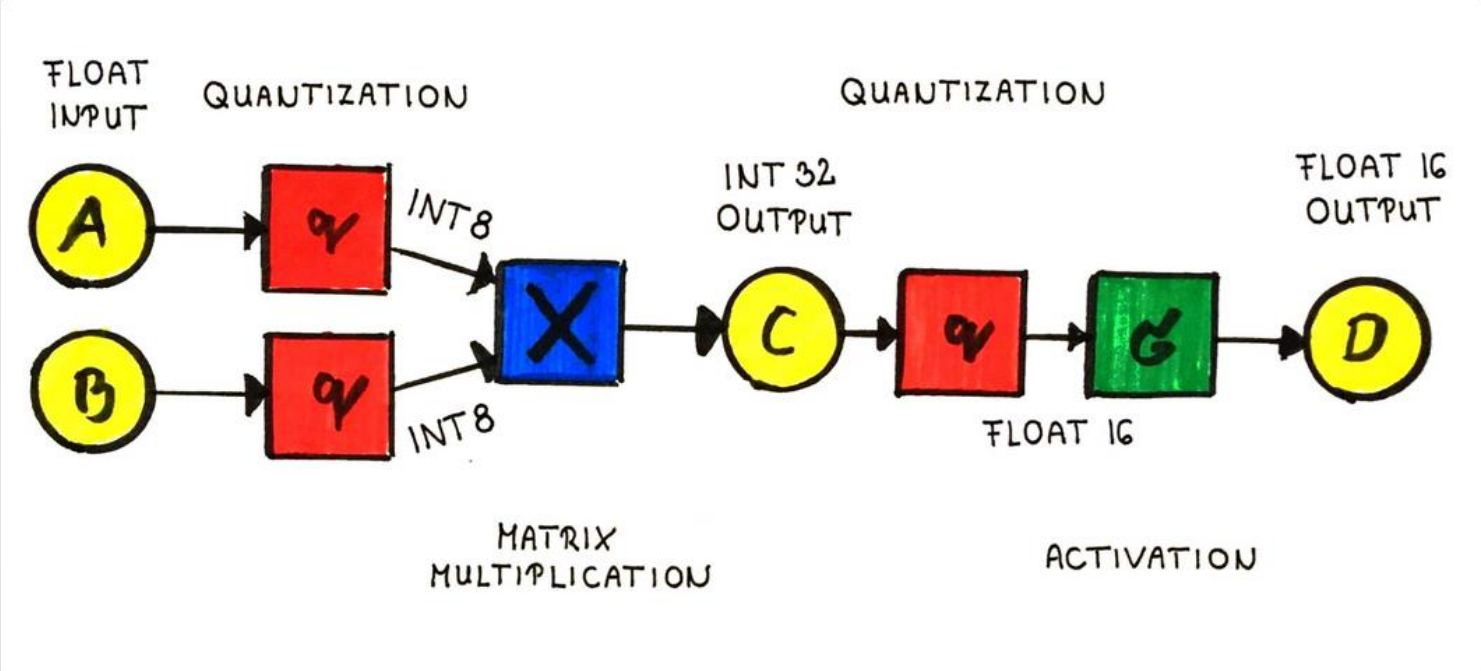
- Sunt gata să pun modelul în producție, însă probabil va rula pe hardware diferit față de cel de antrenare
  - Există **optimizări specifice** pe care le pot face?



1. Colectare
2. Pregătire
3. Antrenare
4. Evaluare
5. **Producție**
6. Re-antrenare

5.1 Quantizare

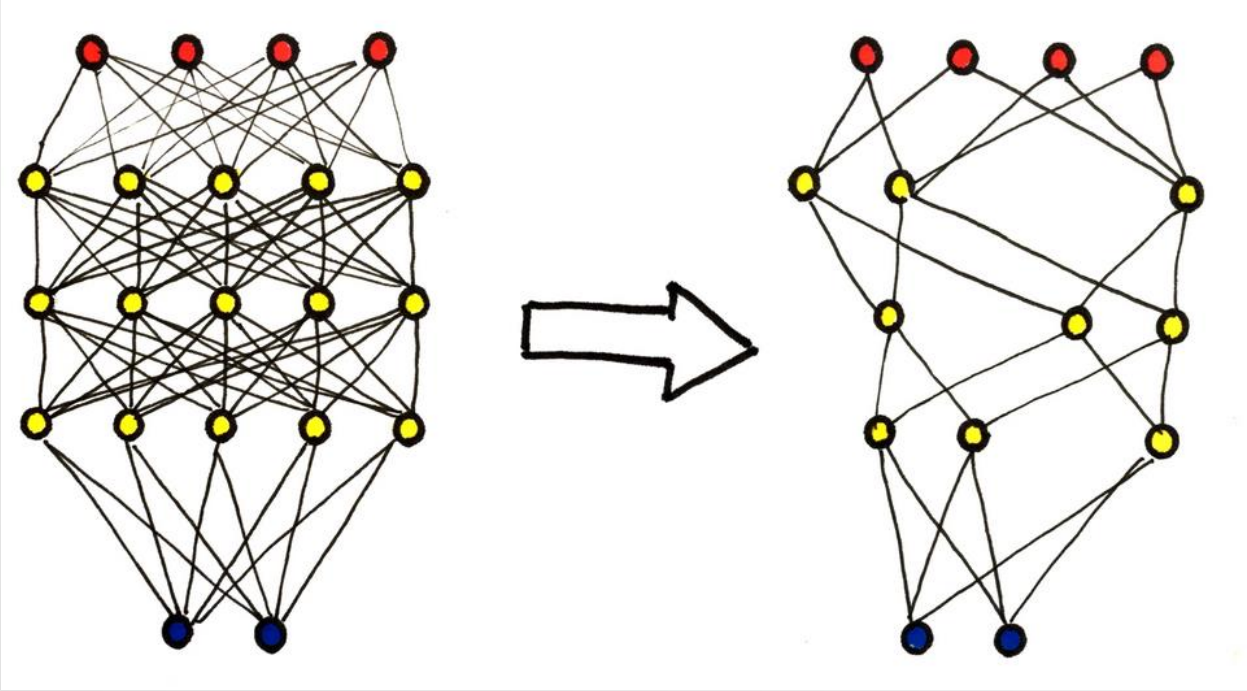
- Reducerea numărului de biți prin care este reprezentat un număr
- Modelul e antrenat pe GPU-uri și folosește floats pentru a calcula parametrii
- În producție, acești parametri pot fi quantizați la int8,
  - Un câștig semnificativ de viteză și de spațiu



Sursă

5.3 Pruning

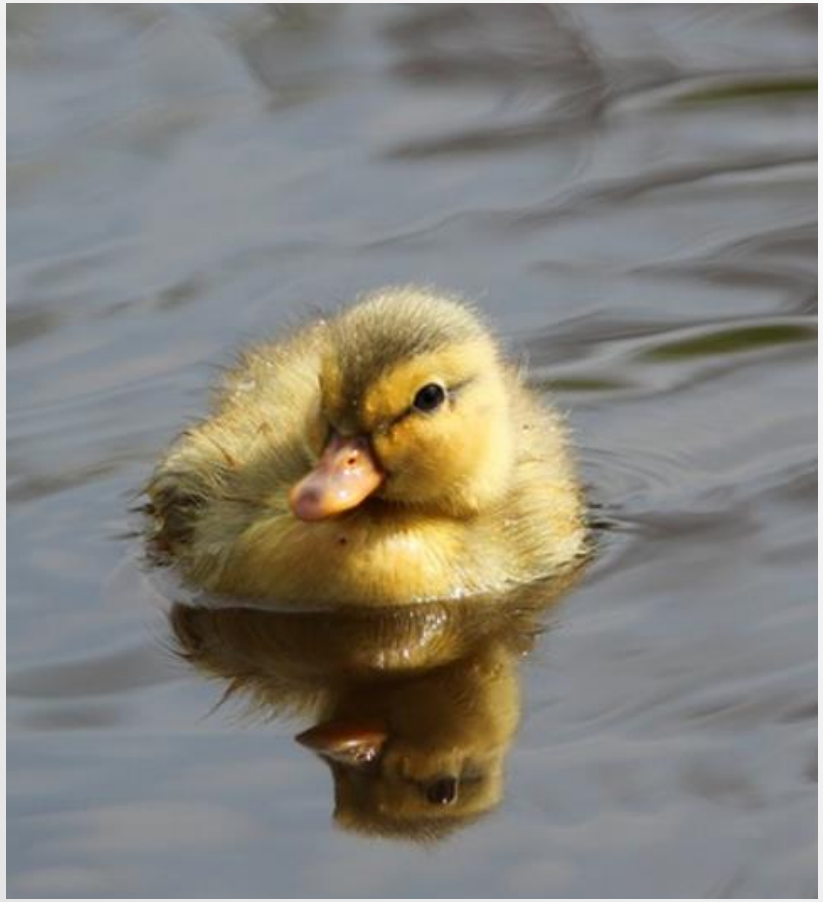
- Sunt sigur că modelul are nevoie de toate căile pentru a face o predicție?
- *Poate pot scoate unele noduri din graful meu computațional, și astfel cresc performanța.*



Sursă

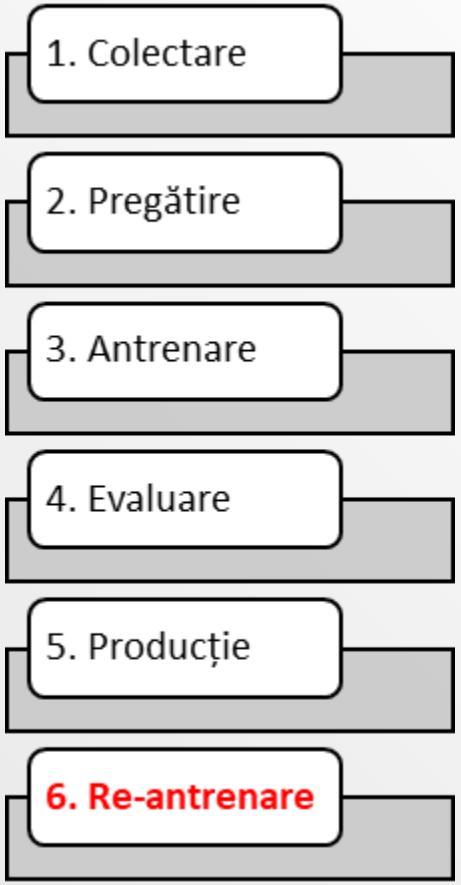
5.4 Re-evaluare și antrenare continuă

- Re-evaluez performanța modelului în '**lumea reală**'.
  - *Evaluările obiective pe un set separat de date sunt cruciale, însă nu pot fi sigur de performanța modelului meu până nu îl arunc în realitate.*
- **Continui** să antrenez modelul cu datele noi.
  - *Antrenarea nu se oprește după producție!*
  - *Procesul este ciclic și continuu.*
  - *Pot adăuga exemplu cu exemplu în modelul meu, uneori chiar în timp ce rulează!*



6. Re-antrenare

- **Reluarea totală** a procesului de învățare
  - Modelul meu **nu mai are performanța așteptată**
  - S-au schimbat foarte mult **condițiile inițiale**
    - *Modelul meu a mers foarte bine vara însă pentru iarnă nu are deloc date*
  - Am strâns o **cantitate foarte de date noi**
    - *Adăugarea incrementală de date în model are rezultate mai slabe decât adăugarea unui batch și reantrenarea.*
    - *Dacă am strâns multe date, ar trebui să reantrenez.*
  - Au apărut **arhitecturi** mai bune
    - *Lumea ML-ului evoluează foarte rapid și pot apărea idei semnificativ mai bune de a rezolva o problemă*

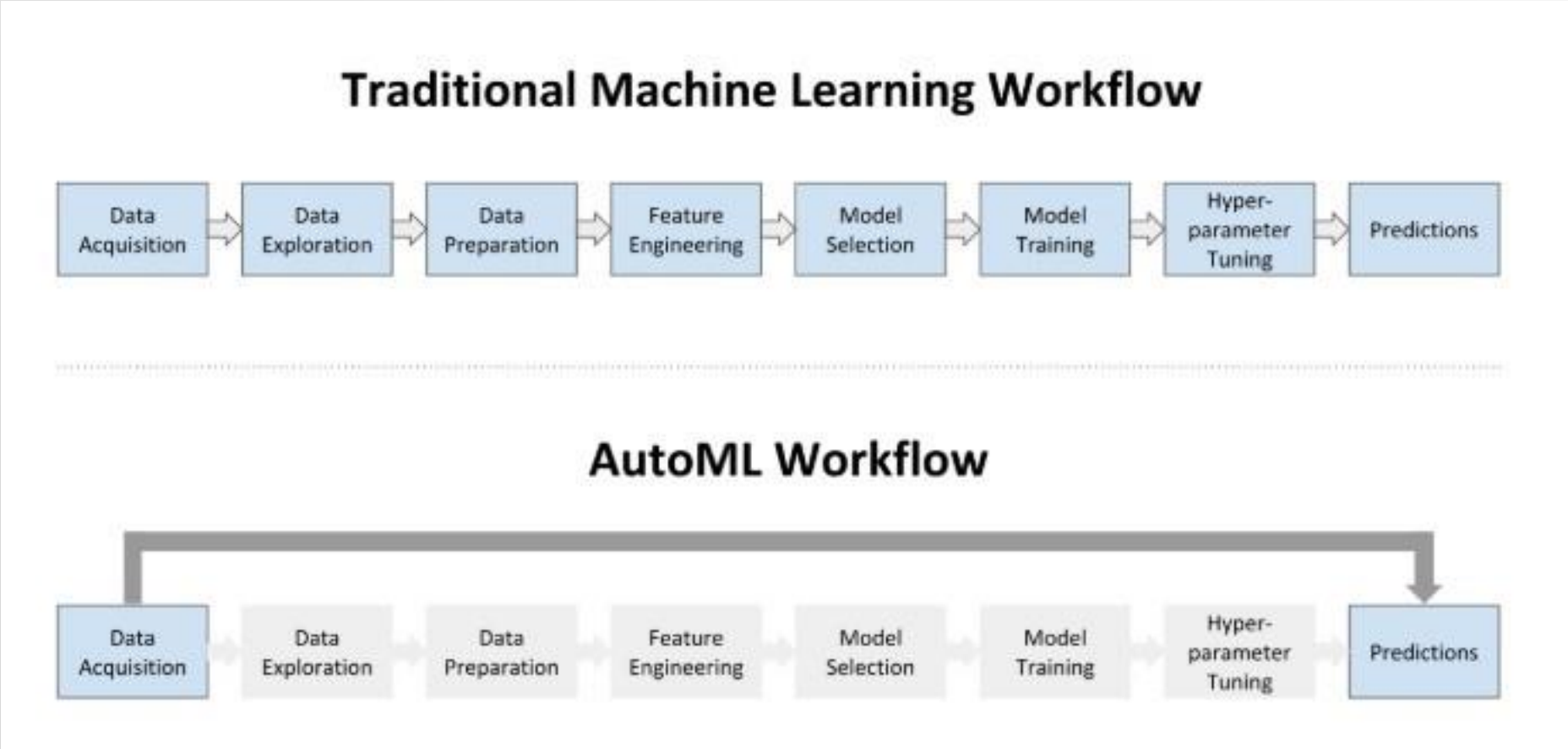


#3.4 Auto-ML

Folosirea ML pentru a construi ML

Auto-ML

- Automatizarea dezvoltării, evaluării și rafinării modelelor



Auto-ML: Input - Output

- Input: setul de date și problema de rezolvat
- Output: rafinarea datelor, pregătirea și specificarea modelului, predicții
- Similar cu identificarea celei mai bune arhitecturi pentru piesele LEGO disponibile

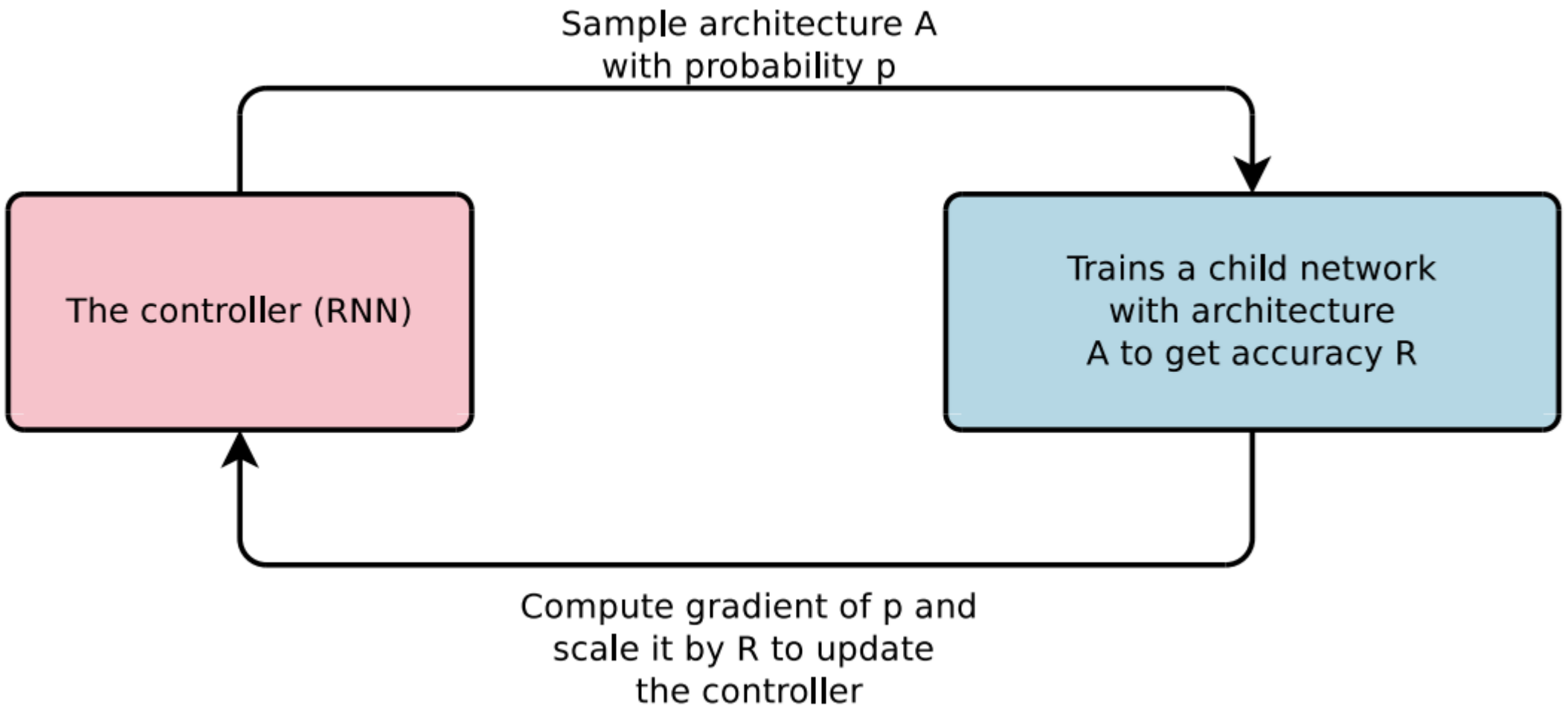
Auto-ML: Dezavantaje

- Este **extrem** de costisitor computațional. Numai giganții industriei își pot permite o implementare, și numai ei oferă soluții.
- Implică **folosirea API-ului** lor pentru predicție
- Nu îți poți vedea modelul creat!
  - Nu există privacy, totul trecând prin cloud
  - Nu poți da deploy de un device offline sau mic, cum ar fi Raspberry Pi
  - Nu poți învăța ce l-a făcut pe model bun, și asta nu îți amortizează din costul de producție pe termen lung

Auto-ML: Limitări

- Cineva tot trebuie să ofere acele 'bucăți de LEGO' inițiale modelului, altfel căutarea ar fi incredibil de mult mai lungă. Iar alegerea lor nu este deloc trivială.
- Formularea problemei încă rămâne în mâinile noastre, și aceasta de multe ori este cel mai important pas.

Funcționarea auto-ML



Pe scurt: procesul ML

3.1 De ce ML?

3.2 Tipuri de probleme pentru ML

A. Învățarea supravegheată

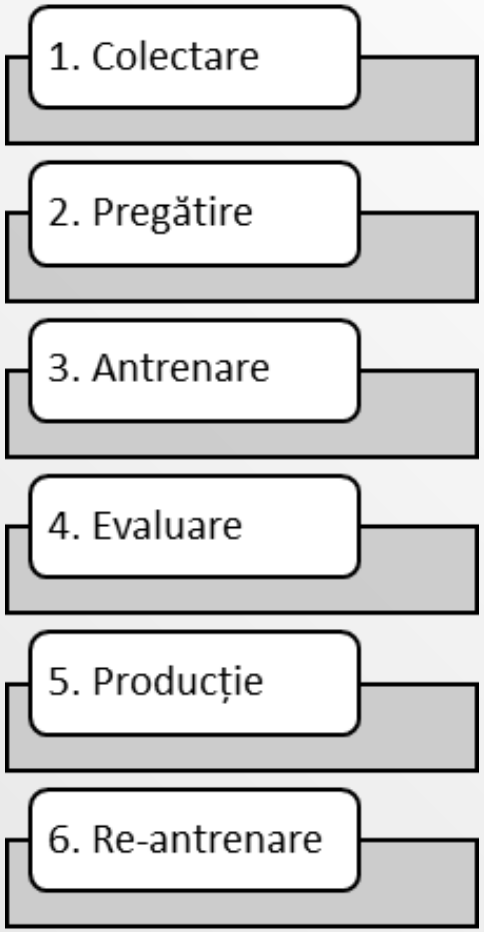
1. Clasificare
2. Regresie & serii de timp
3. Sequence to Sequence

B. Învățarea nesupravegheată

1. Clustering
2. Reducerea dimensionalității
3. Transfer learning
4. Reinforcement learning

3.3 Procesul învățării automate

3.4 Auto ML



Referințe

1. Zinkevich, M. Best practices for ML engineering. Google
2. MIT Technology Review (2019). A US government study confirms most face recognition systems are racist.
3. Poyser, O. (2019). Exploring the dynamics of Bitcoin's price: a Bayesian structural time series approach. Eurasian Economic Review, 9(1), 29-60.
4. Storcheus, D. (2016). Understanding dimensionality reduction and its applications
5. Heaven, W.D. (2020). The way we train AI is fundamentally flawed. MIT Technology Review

SYSTEMS LABORATORY

Cursul DS04

Tehnologiile datelor: Machine Learning și Deep Learning

PR network crunch

Structura cursului

- Introducere: De ce DS?
- Scopul cursului
- Măsurarea și colectarea datelor
- Erori de măsurare
  - Precizia
  - Validitatea
  - Fidelitatea
- Corelația
- Erori de modelare
  - Corelații iluzorii
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erori de eșantionare
- Tehnologiile datelor: programare vs. Învățare automată
- Procesul învățării automate
- Machine Learning și Deep Learning
- Predicție și cunoaștere: analiza datelor
  - Măsurare și reducerea dimensionalității
  - Clasificare (cluster)
  - Cauzalitate și regresie
  - Predicții și serii de timp
- Predicția ca premisă pentru acțiune
- Predicția ca produs economic

DS01: Pe scurt despre știința datelor

- Abilități esențiale privind lucrul cu informații
- Abilități tot mai cerute datorită avansului AI și Big Data
- Două competențe cheie: să ții cont simultan de...
  - Procesul măsurat și procesul măsurării
  - Tendința centrală, variabilitate și excepții
- Tendința centrală e estimată prin: medie, mediană, mod
- Variabilitatea: variația, deviația standard
- Erorile de măsurare pot privi:
  - Precizia: sistematice / aleatoare
  - Validitatea
  - Fidelitatea (repetabilitatea)

DS02: Pe scurt despre corelație și erori

- Corelația este un posibil simptom al cauzalității
  - Corelația este observabilă în norul de puncte
  - Coeficienții de asociere măsoară intensitatea relației
- Modelul face legătura dintre corelație și explicație
- Erori de modelare
  - Sensul cauzării
  - A fi corect din rațiuni false
  - Corelații coincidentale
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erorile pot proveni nu doar din model, ci și din eșantionare

DS03: Pe scurt despre procesul ML

3.1 De ce ML?

3.2 Tipuri de probleme pentru ML

A. Învățarea supravegheată

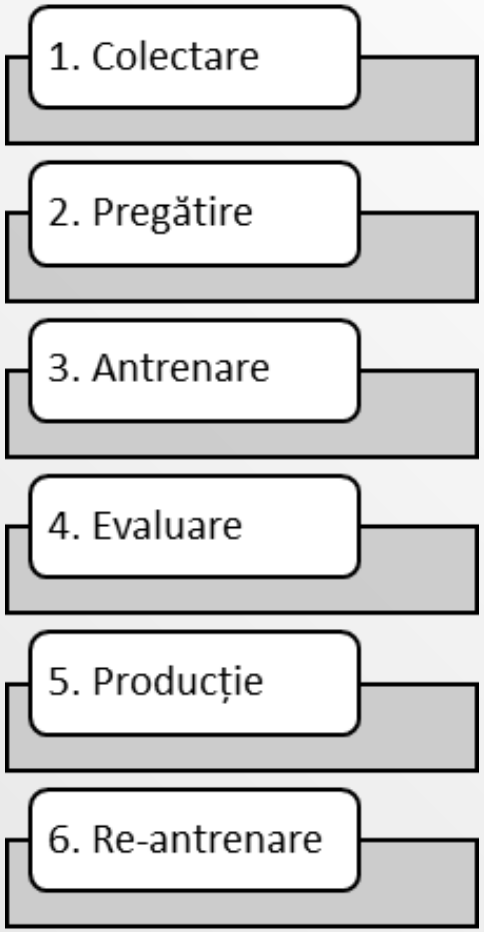
1. Clasificare
2. Regresie & serii de timp
3. Sequence to Sequence

B. Învățarea nesupravegheată

1. Clustering
2. Reducerea dimensionalității
3. Transfer learning
4. Reinforcement learning

3.3 Procesul învățării automate

3.4 Auto ML



Bleeding-edge AI

- Timnit Gebru et al., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”
- 4 riscuri ale modelelor large-scale AI (antrenate pe Vast Data)
  - Carbon footprint & costuri financiare
  - Modele „black box”: rasism, sexism, bias
  - Cost de oportunitate pentru modele mai „inteligente” antrenate pe date mici
  - Iluzia sensului: generarea fake content

Timnit Gebru

Computer scientist

ai.stanford.edu/~tgebru

Timnit Gebru, is a computer scientist who works on algorithmic bias and data mining. She is an advocate for diversity in technology and co-founder of Black in AI, a community of black researchers working in artificial intelligence. Wikipedia

Born: Ethiopia

Nationality: Ethiopian

Research interests: Fairness

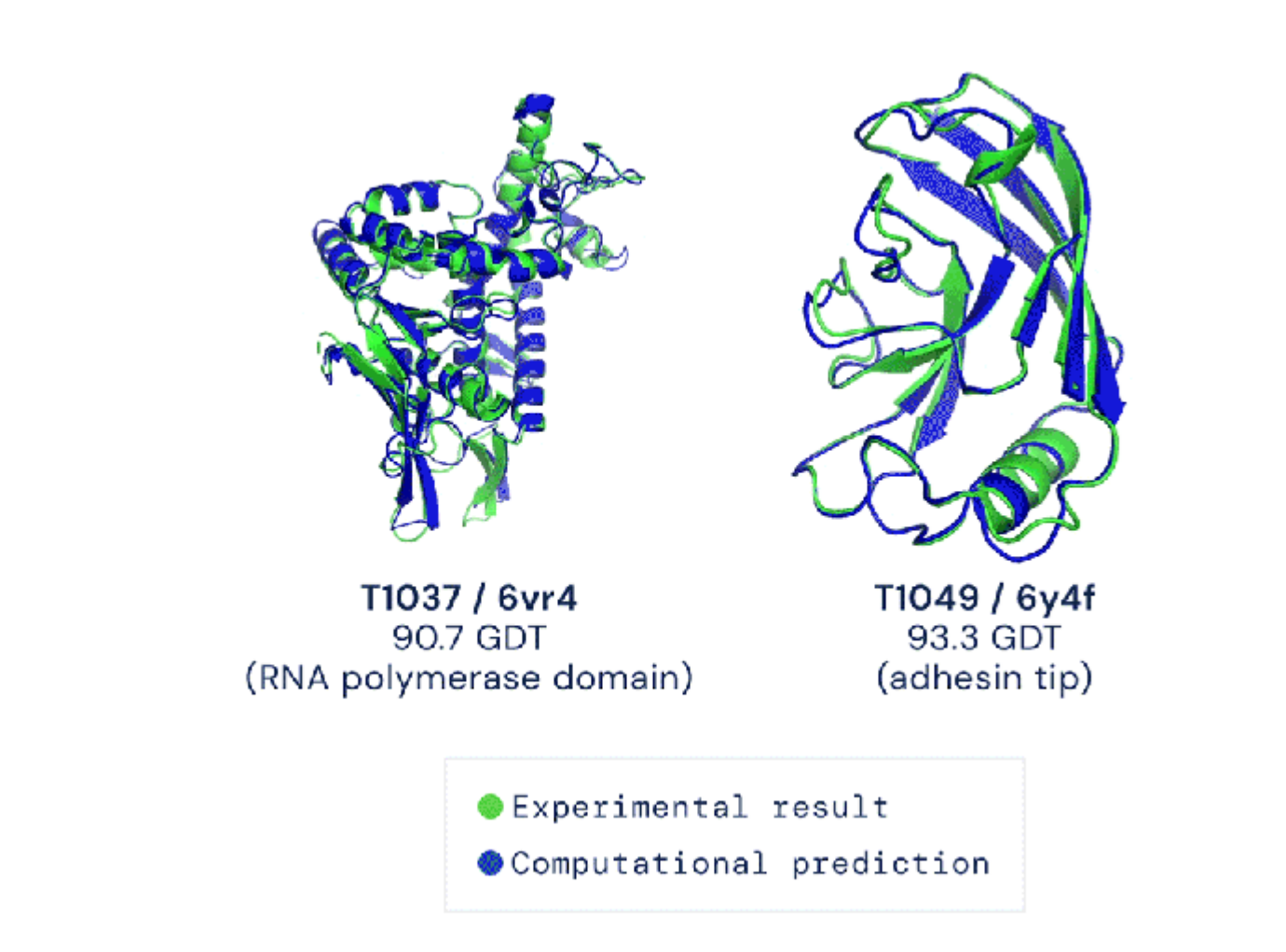
Education: Stanford University

Academic advisor: Fei-Fei Li

Alphafold 2

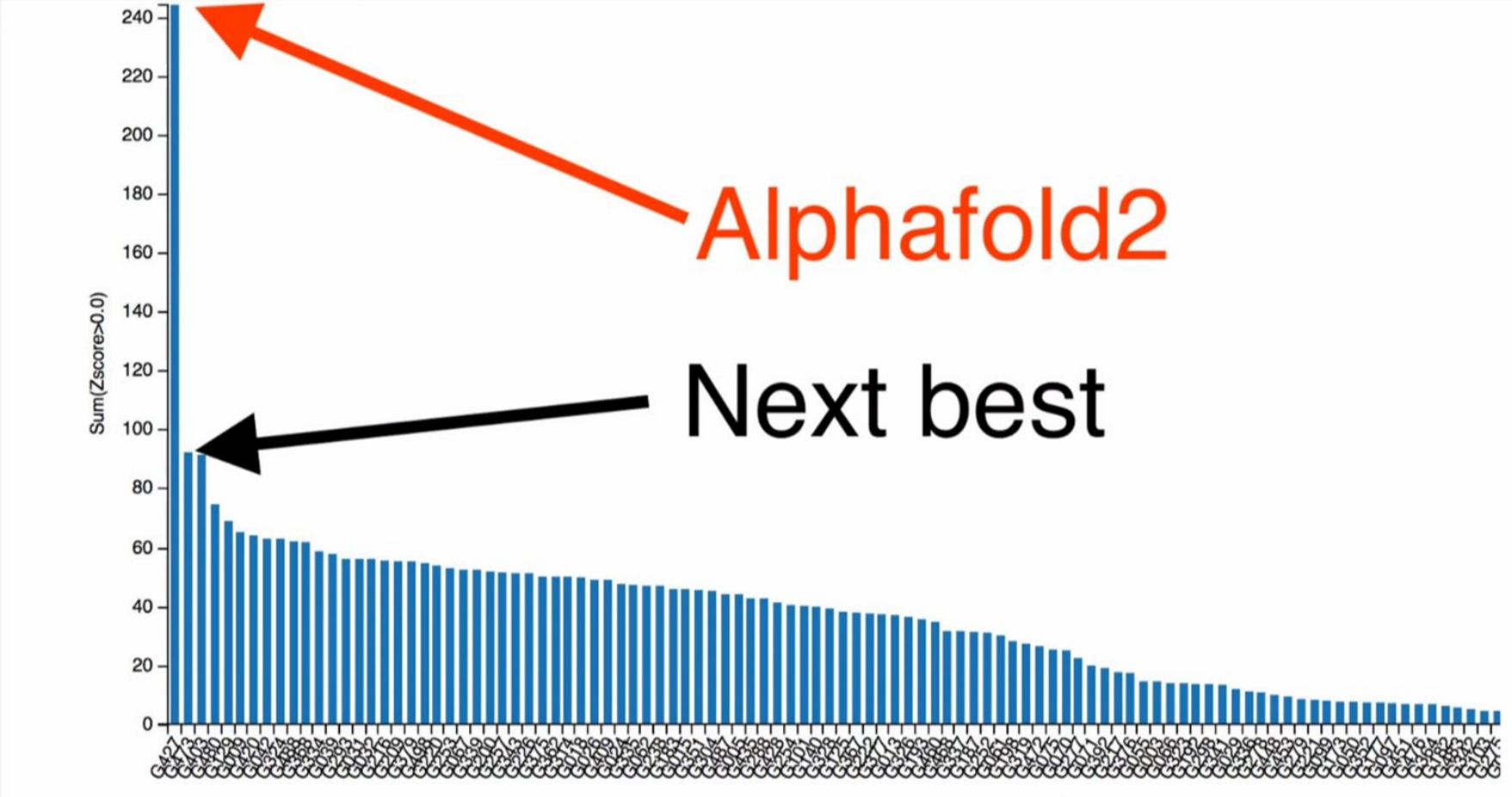
- 30 Nov 2020, DeepMind a rezolvat problema de protein folding. [Wiki]

- Problema a rămas aproape neatinsă în ultimi 50 de ani.



Mecanisme de Atenție

- Alphafold2 a depășit substanțial competiția de anul acesta.
- Pare că s-au folosit de transformers pentru a aceste rezultate



Cuprins

4.1 Procesarea vocii

- Audrey: problemele predicției modulare
- Rețele neurale: avantaje și probleme

4.2 Procesarea Limbajului Natural (NLP)

- De la Bag of Words la Transformers

4.3 Procesarea Imaginilor

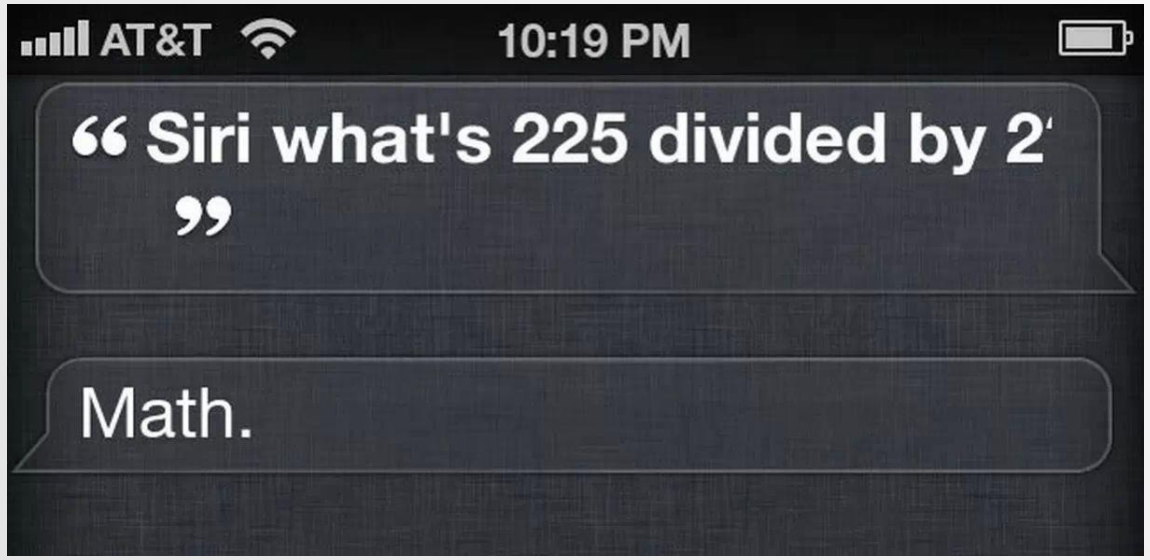
- Convoluted Neural Networks (CNNs)
- Transformers pentru imagini?

#4.1 Procesarea vocii

Voice recognition: De la Audrey la rețele neurale

Procesare audio

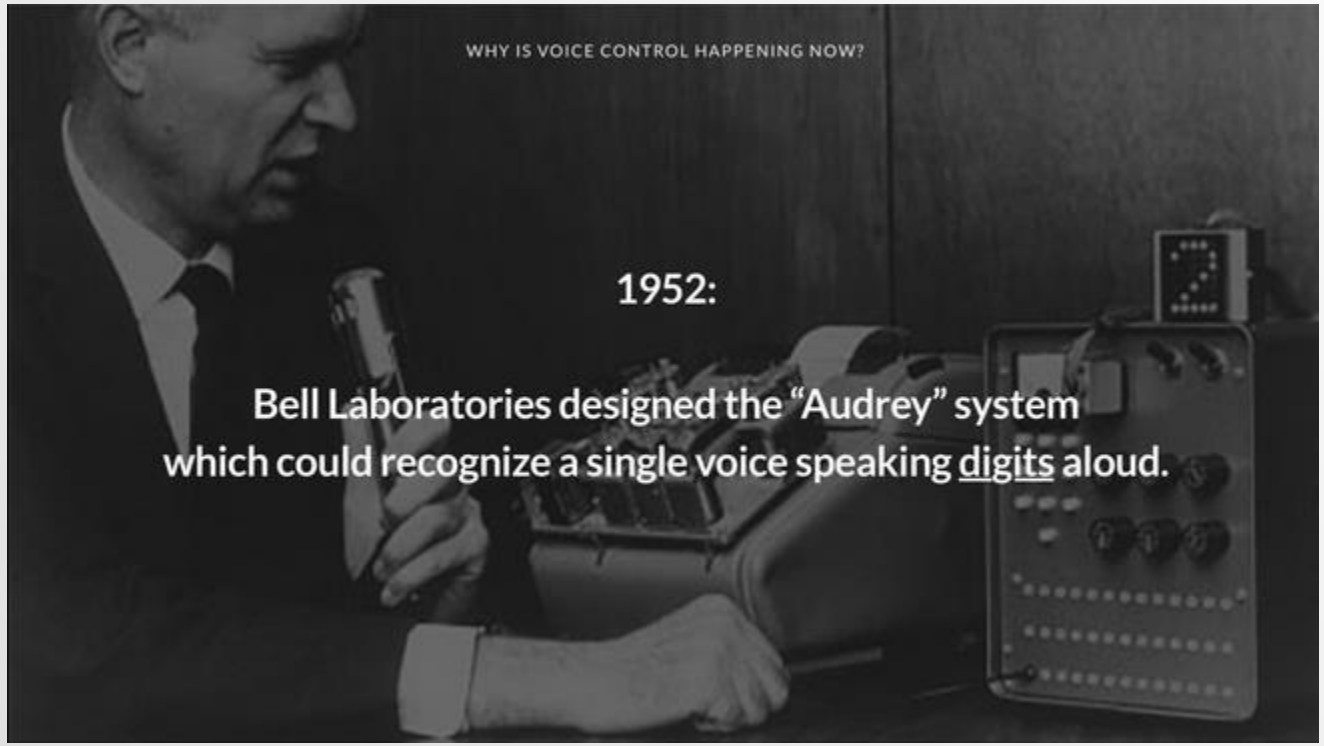
- Sistemele automate pentru înțelegerea vocii umane au evoluat incredibil în ultimele decenii.
- Ele erau extrem de limitate până recent, algoritmii fiind atât de inconsistenți încât cu greu își câștigau numele de euristică.



Sursă

De unde am pornit? Sistemul Audrey

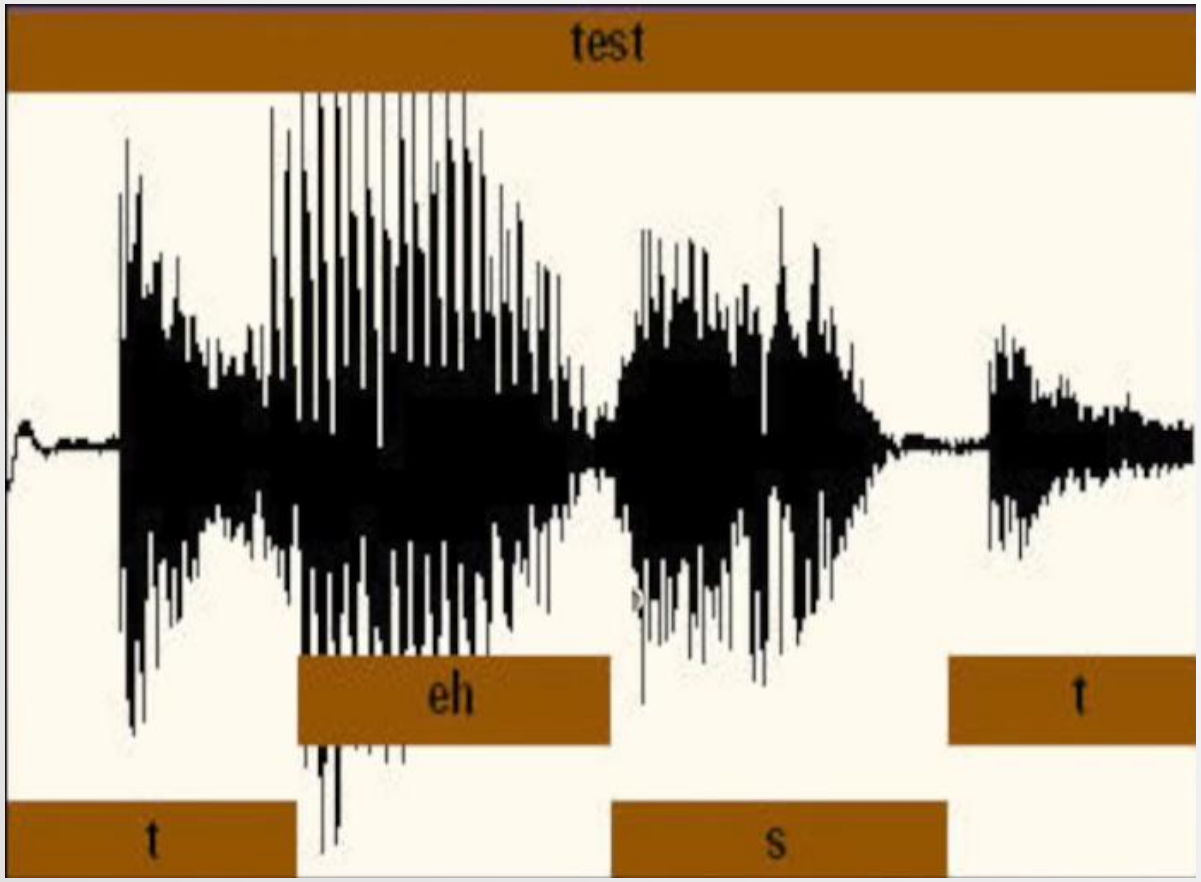
- Primul "sistem" de speech recognition a apărut în 1952: Audrey (Bell Laboratories)
- Putea identifica doar cifrele de la 0 la 9
- Pentru o singură voce



Sursă

Sistemul Audrey | 2

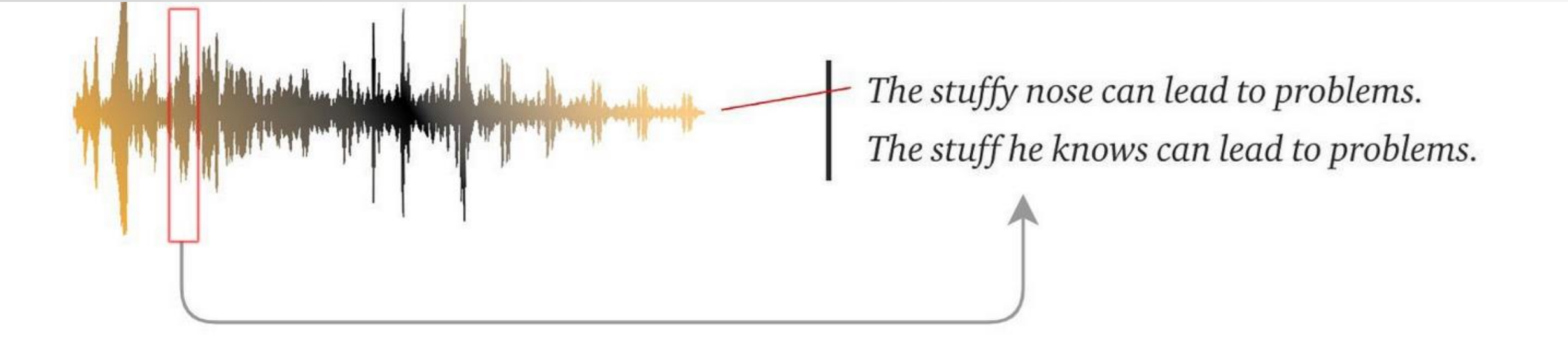
- Împărțea un semnal audio în unități fonetice, care apoi pot fi identificate cu ajutorul unor algoritmi, mapate și în final reconstruite.
- Rezultate foarte slabe!
- Acest sistem a fost îmbunătățit, ajungând la vocabulare de zeci de mii de cuvinte
- Până la apariția rețelelor neurale, problema nu a fost nici pe aproape rezolvată.



Sursă

Prezicere modulară: Care e problema?

- Unitățile fonetice pot diferi foarte mult
  - Cuvintele pot fi spuse mai rapid sau mai alungit,
  - Pentru noi este același cuvânt, însă semnalul audio arată drastic diferit
- Prezicere modulară este foarte limitată
  - Diverse litere sună foarte diferit în diferite cuvinte / contexte de propoziție



Prezicere modulară | 2

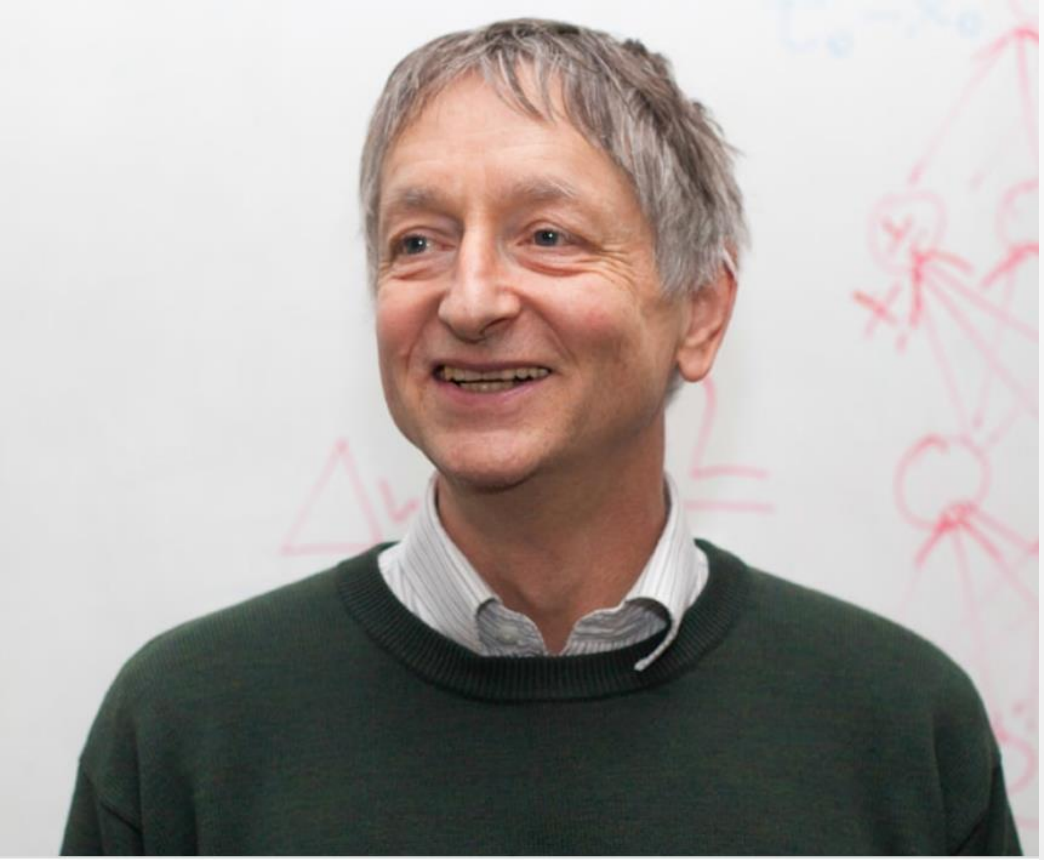
- Sistemul este extrem de sensibil la zgomote.
  - Orice zgomot de fundal poluează spectrograma sunetului
  - Depărtarea de microfon contează foarte mult, adăugând ecou
- Nu există generalizare
  - Unui astfel de sistem îi este foarte greu să se adapteze peste diferite accente sau chiar limbi

Prezicere modulară | 3

- Performanță foarte proastă cu mai mulți interlocutori.
  - Dacă nu avem fiecare vorbitor pe un canal diferit, un astfel de sistem nu are nici o șansă să diferențieze în mod consistent un număr mai mare de persoane într-o convorbire.
- Contextul nu este luat în calcul.
  - Noi interpretăm cuvintele și propozițiile cu totul, unele cuvinte pot suna foarte ciudat în afara contextului.

Tranziția spre rețele neurale

- Multe alte idei s-au încercat
  - IBM a propus un sistem bazat pe un Hidden Markov Model care părea promițător
  - RN ofereau rezultate proaste
- Geoffrey Hinton, un pioner al învățării automate, a insistat în ciuda descurajărilor
- Nimeni nu îl credea până când... RN au început să mai bune decât orice altceva.



Geoffrey Hinton, Wiki

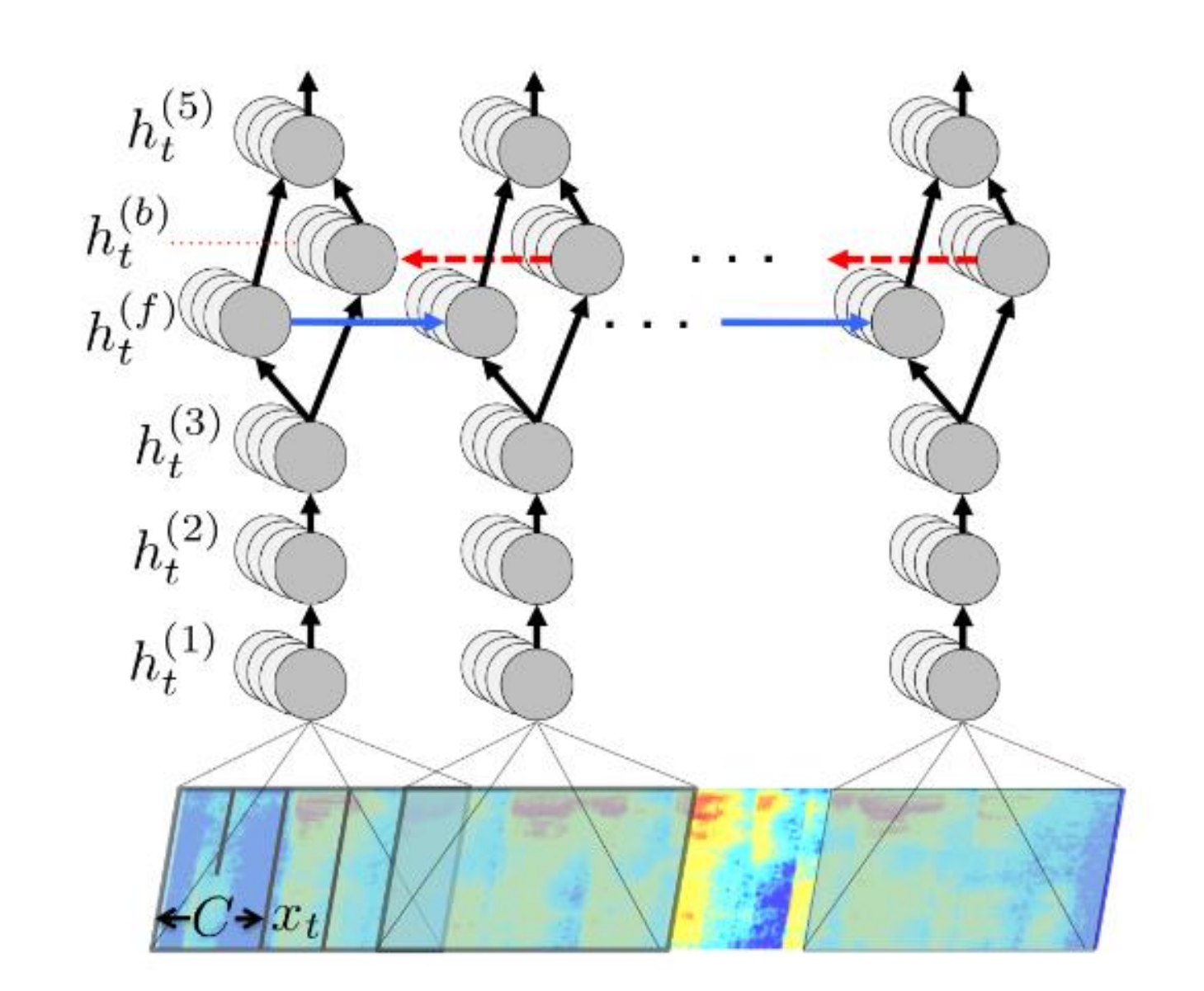
Rețele neurale: De ce?

- Rețelele neurale se folosesc din plin de:
  - Cantitatea imensă de date din ultimii ani
  - Capacitatea de paralelizare a noilor plăci video
- Ce date avem?
  - Nu avem fișiere audio în care fiecare literă este etichetată...
  - Avem filme subtitrate!
  - Avem YouTube
  - Mozilla Voice: zeci de mii de ore de exemple

Rețele neurale | 2

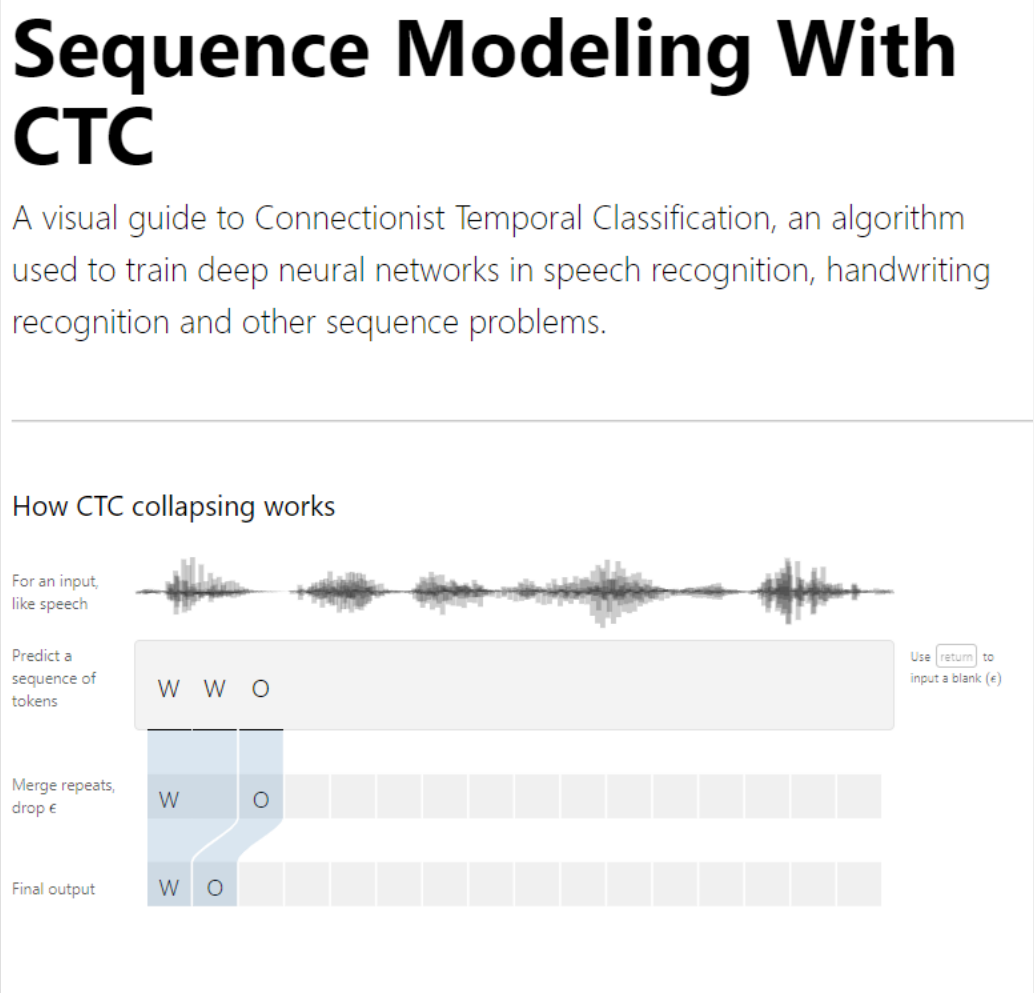
- Modelăm predicțiile ca un vector de probabilități
  - Evaluăm contextual în ambele direcții
  - Modelul își actualizează predicțiile cu date noi în context
- Arhitecturi recurente: se uită în trecut și în viitor pentru decizia din prezent
  - Coerență temporală
  - Paralelizarea antrenărilor

Rețele neurale: Arhitecturi recurente



Rețele neurale: testare

- Folosim algoritmi agnostici la aliniament pentru a evalua performanța modelului
  - De ex: CTC – Connectionist Temporal Classification



Rețele neurale: generare cazuri

- Sporirea setului de date
  - Zgomot de fundal: putem pur și simplu să-l adăugăm peste datele pe care le avem deja
  - Pitch-ul și ecoul pot fi artificial adăugate într-un mod realist

Recunoașterea vocii: Privacy

- Aproape toate soluțiile curente se folosesc de cloud pentru computație
  - Vocile noastre ajung pe serverele companiilor
  - Există modele lightweight care pot rula local: Mozilla DeepSpeech (excepții)
- Vocile = identificatori biometrici
  - Putem fi identificați prin date viitoare CCTV sau convorbiri telefonice etc (The Guardian 2019)
  - „Function creep”: utilizarea informației în afara contextului inițial

#4.2 Procesarea limbajului natural NLP

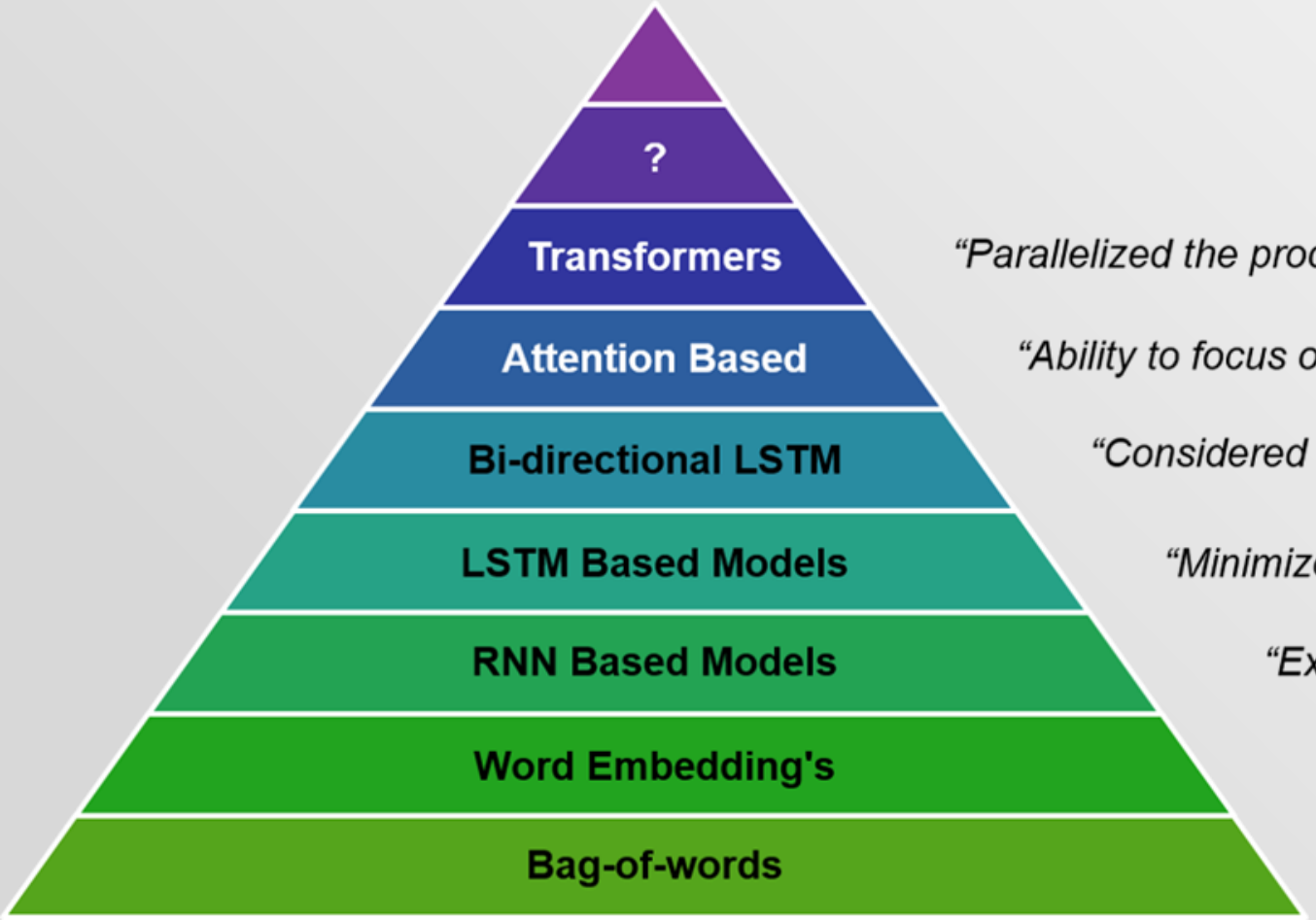
Procesarea Limbajului Natural (NLP)

- Inițiată în anii 1950: Testul Turing (Wiki)
- Inițial: bazată pe reguli complexe 'scrise de mână', pe arbori de decizie și pe modele statistice
  - Astfel de modelări nu pot cuprinde complexitatea limbajelor și nuanțele lor
  - Prea multe reguli!
- Actualmente: bazate pe ML
  - ML e excelentă pentru sisteme cu prea multe reguli

Alan Turing, 1912-1954

Reprezentarea numerică a cuvintelor

- **Word Embeddings**
  - Pentru a ne putea aplica algebra lineară peste intrările noastre, cuvintele trebuie reprezentate numeric.
  - Există diverse metode de a transpune cuvinte într-un spațiu numeric



"Parallelized the processing of sequential data, for better performance"

"Ability to focus on relevant input via attention weights"

"Considered the complete sentence in order to predict"

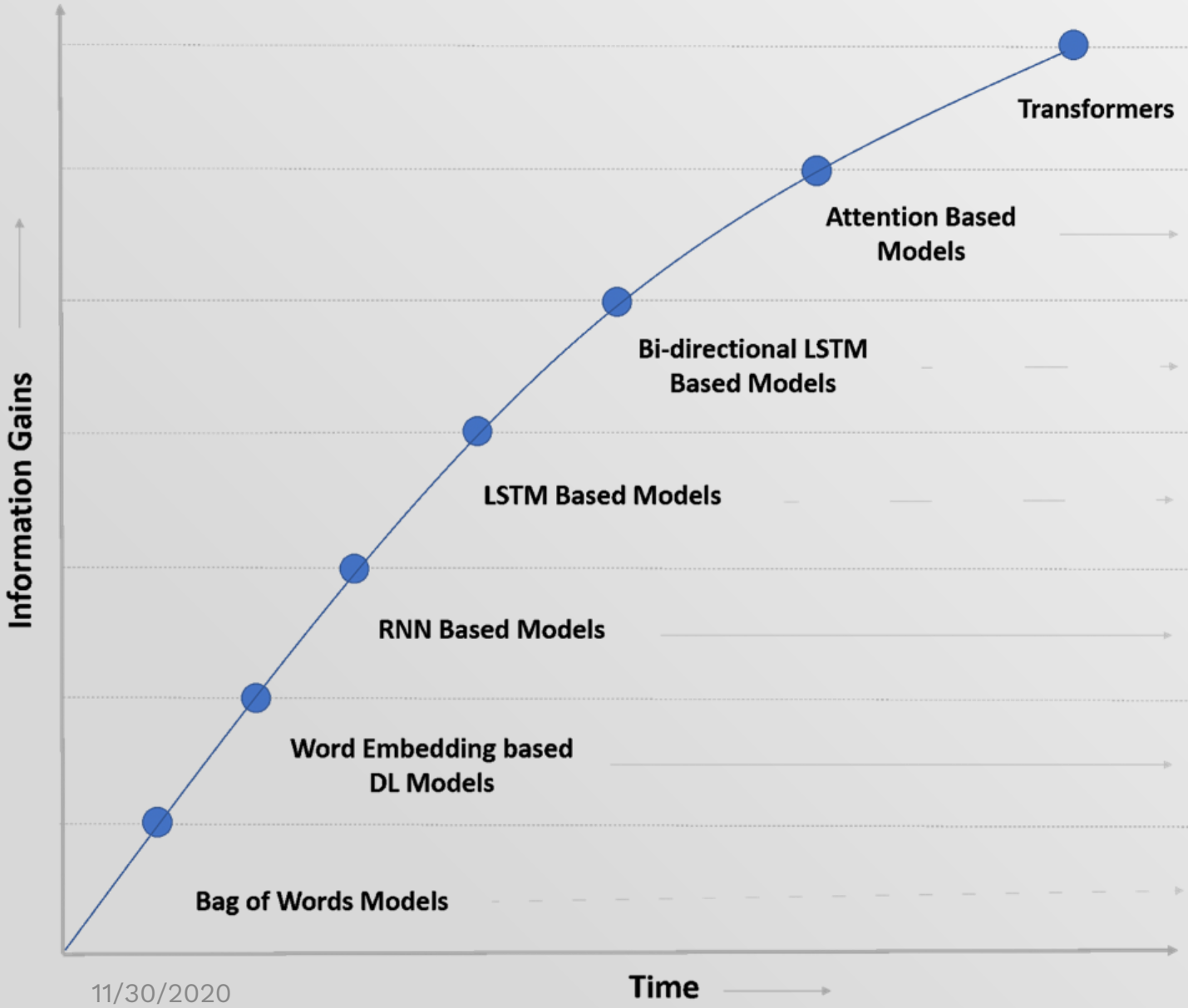
"Minimized the problem of diminishing gradients"

"Extracted positional information"

"Extracted meaning or intuitions from Represented words"

"Converted textual information to numerical form"

Source

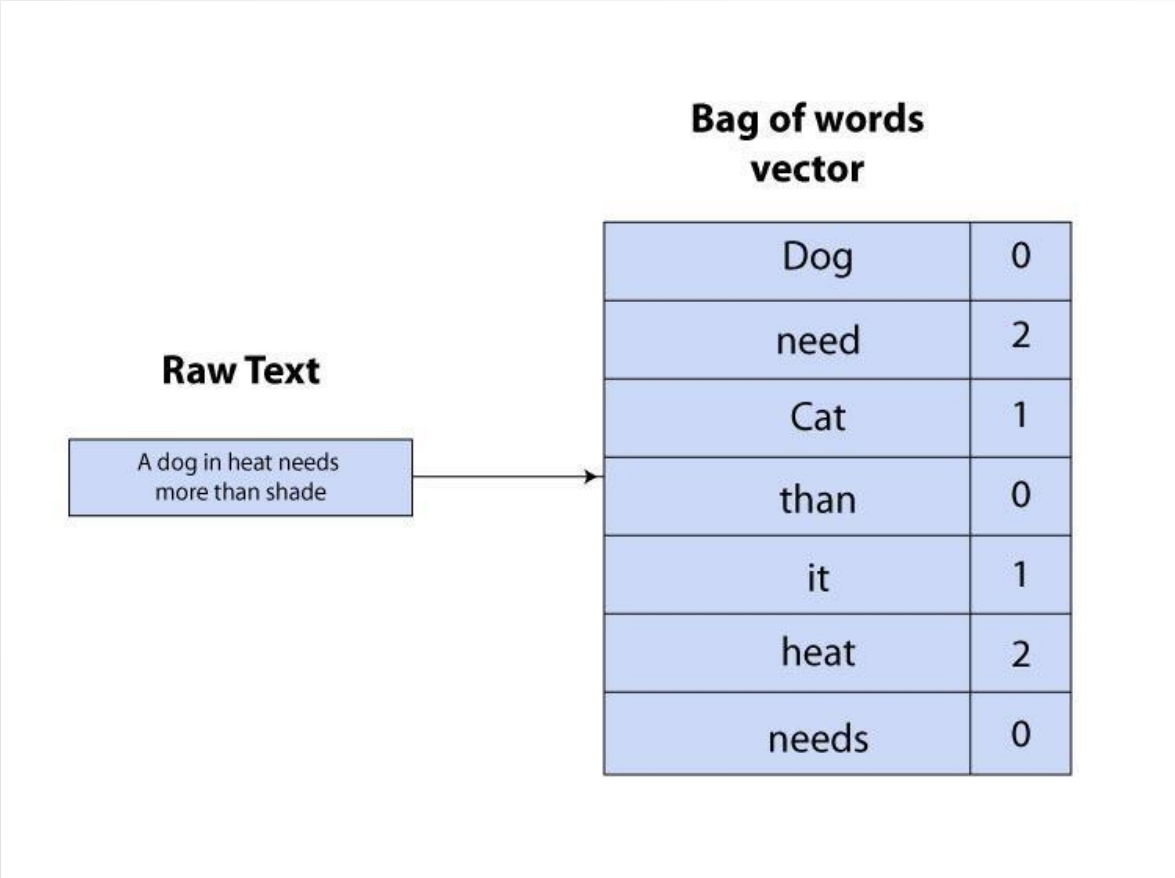


Advancements Made	Challenges that still persisted
Built on Attention mechanism using encoder-decoder. Helps parallelization of sequential data	Doesn't make effective use of context information, sematic info not respected. Transformer XL addresses these issues
Model has mechanism in place to focus attention+ of relevant pieces of input	Computation was done sequentially. Was difficult to scale up.
Models built based on both past and future data	Need complete sequence of data to predict. Compressing complete input source leads to dip in performance
Mechanism in place to Update/Forget relevant information deep into network	Prediction based on only past data
Models included the positional information apart from the meaning of the word	Model tended to forget the values of earlier inputs, especially in long sentences
Gave more importance to the words and the meaning behind those words	Didn't give importance to the position of the words used
Helped convert textual information into a vector format that can be processed by models to gain insights	Did not attempt to give any importance to the meanings behinds the words used

Source

Bag of words

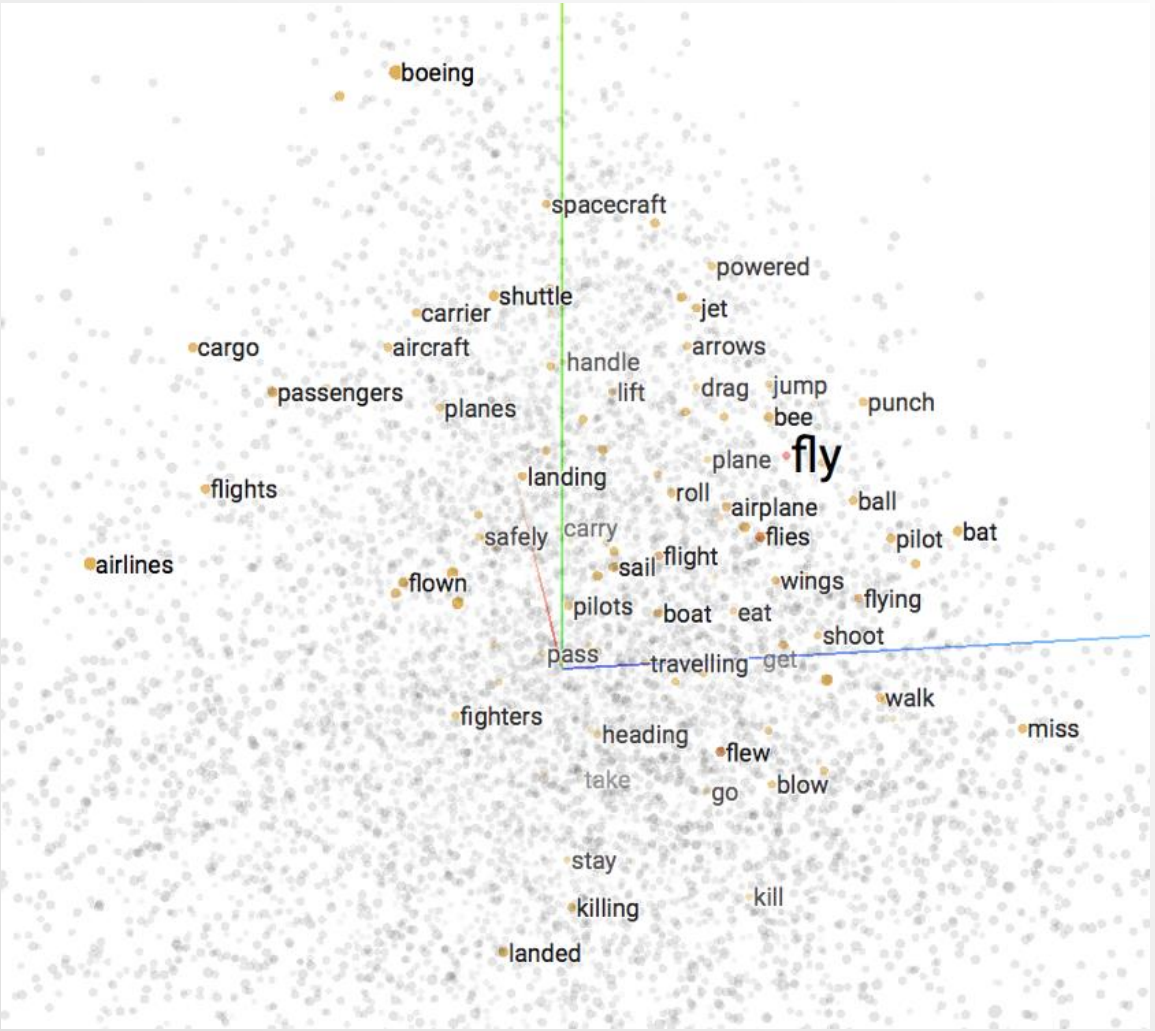
- Reprezentare simplistă
- Utilă pentru clasificarea documentelor
- Ignoră poziția
- Păstrează multiplicitatea (frecvența cuvintelor)



Sursă

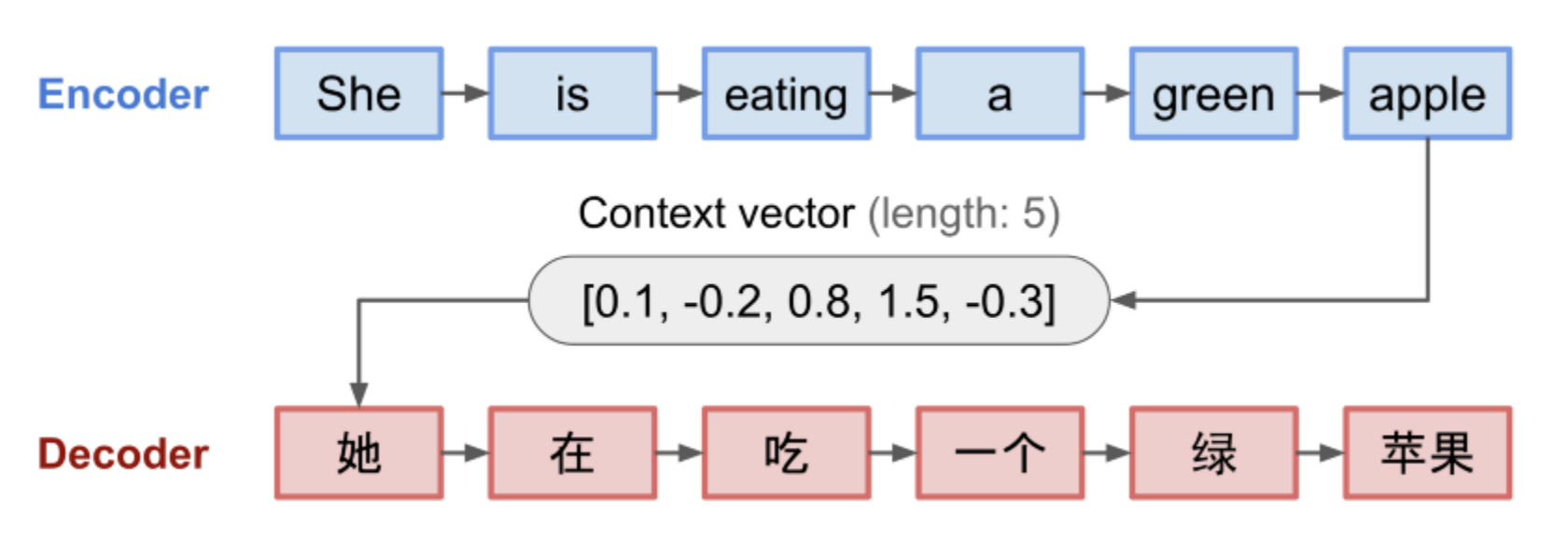
Word2Vec

- Un alt algoritm de encodare a cuvintelor într-un spațiu vectorial
- Problema: este independent de context
  - Același cuvânt are mereu aceeași reprezentare vectorială, indiferent de context
  - Este ignorat polisemantismul



Rețele Neurale Recurente RNN

- Procesează cuvintele secvențial
- Sunt formate din:
  - **Encoder**: cel care creează o reprezentare numerică a intrărilor
  - **Decoder**: cel care transformă acea reprezentare în rezultatul dorit



Rețele Neurale Recurente | 2

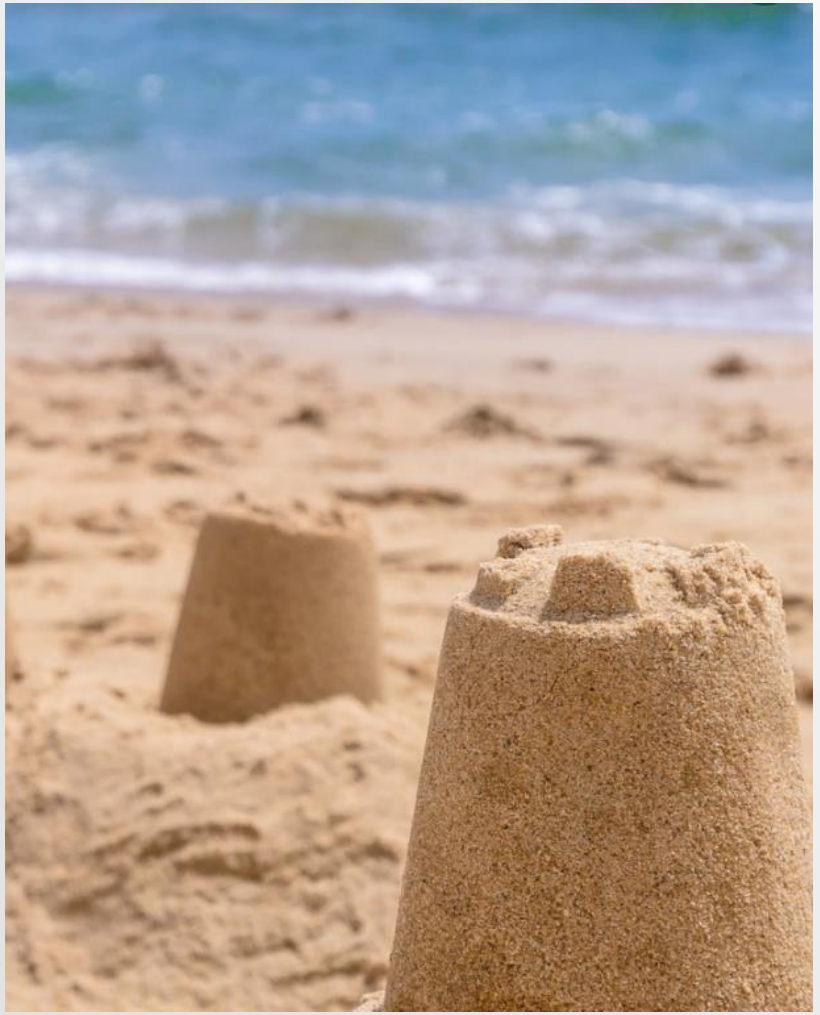
- Pentru a încerca să păstreze contextul, RNN-urile înmulțesc reprezentările vectoriale ale cuvintelor pentru fiecare cuvânt din document, ducând la dispariția sau explodarea gradienților în timpul antrenării

Python

```python
# Pentru un text de 100 de cuvinte
0.9 ** 100 << 1.1 ** 100
```

Long Short-Term Memory (LSTMs)

- Rezolvă unele din neajunsurile RNN-urilor
- vin cu o memorie internă care poate păstra informații anterioare pe care le folosește în decizia curentă.
- Rezolvă și explozia gradienților: adună reprezentările vectoriale la fiecare pas, în loc să le înmulțească.



LSTM | 2

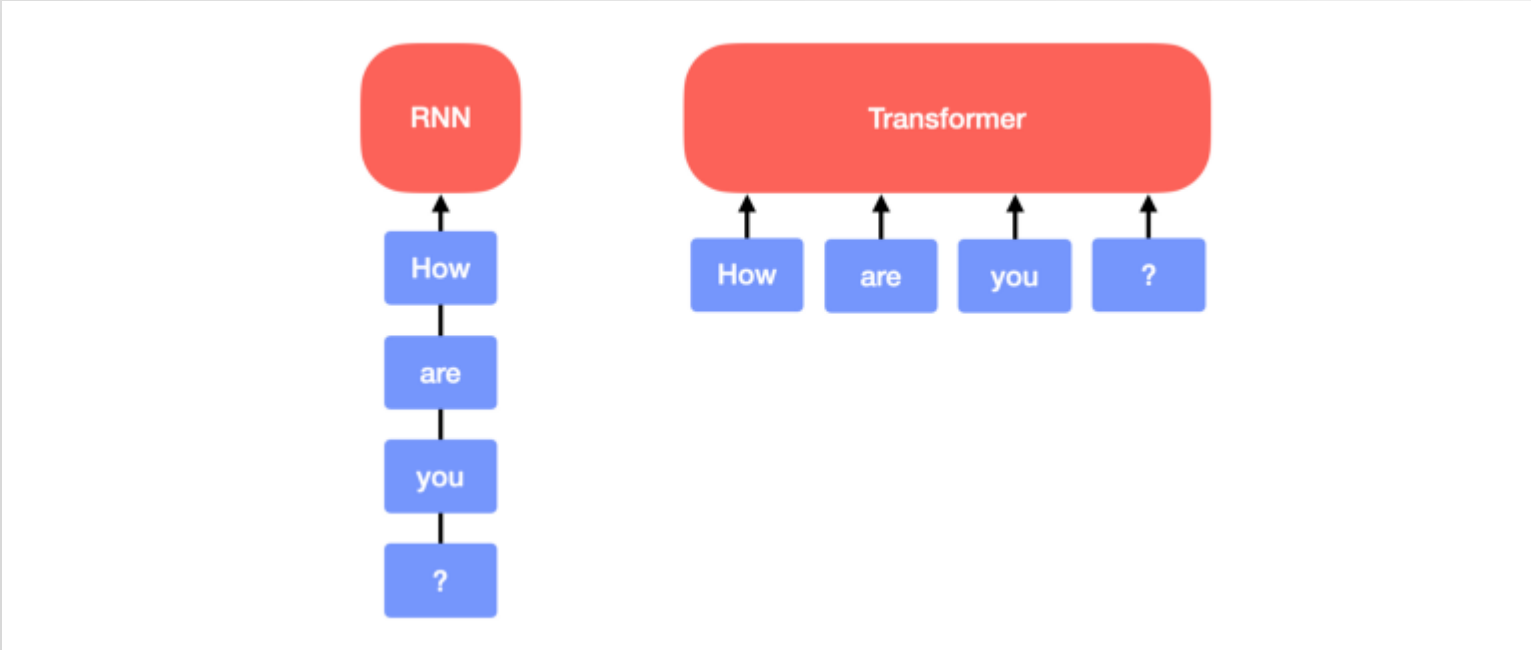
- **Nu sunt paralelizabile**
  - Pentru a păstra un mecanism de memorie, procesările trebuie făcute secvențial
  - Duce la utilizarea foarte ineficientă a hardware-ului și un timp de antrenare mult prea lung
- **Transfer learning-ul nu funcționează**
  - Encodarea datelor în LSTM-uri nu permite transferul înțelegerii pe alte probleme
  - Pentru fiecare nou task, trebuie să reantrenăm un model de la 0
  - Necesită și un set de date adnotat pentru fiecare problemă

Transformers

- Inovație: Vaswani et al, 2017, [Attention is all you need]
- O arhitectură sequence-to-sequence
  - Primește ca și intrare o secvență de cuvinte
  - Toate ca output tot o secvență
  - Propusă pentru problema traducerii

Transformers: atenția

- Mecanism inovativ al atenției
- Rețin doar cuvintele importante din propozițiile anterioare
  - Față de LSTM-uri sau RNN-uri, păstrăm contextul mult mai mult timp, deoarece ținem minte doar cuvintele *importante*



Transformers: paralelizarea

- Primesc ca input cuvintele + pozițiile lor în propoziție
- Permite paralelizarea arhitecturii
  - Eliminăm procesarea secvențială, dar păstrăm ordinea propozițiilor prin aceste embedding-uri de poziție
  - Procesăm simultan întregul input, folosind la maxim hardware-ul

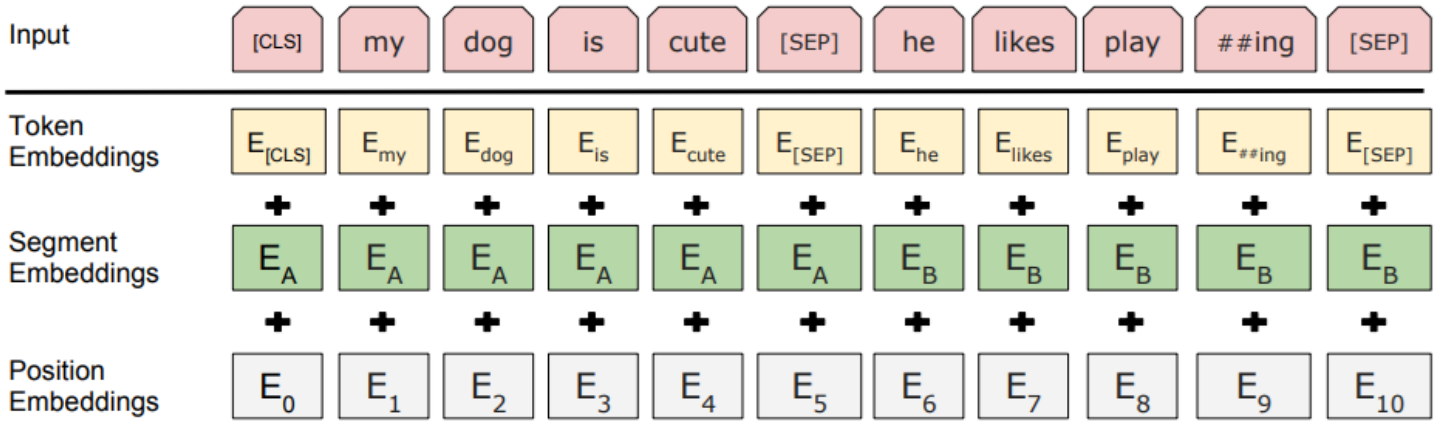
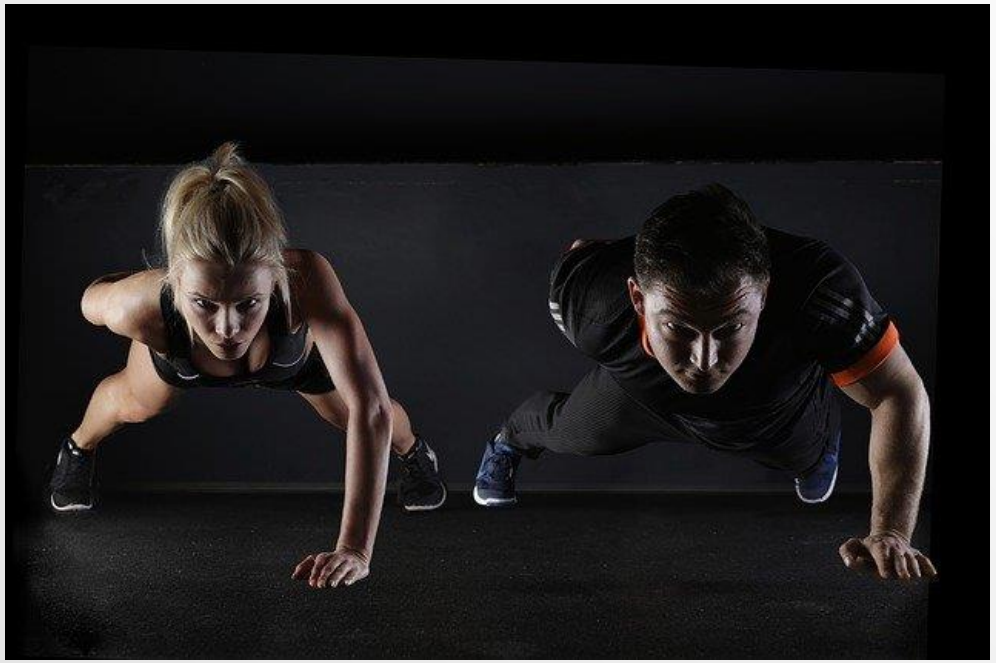


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

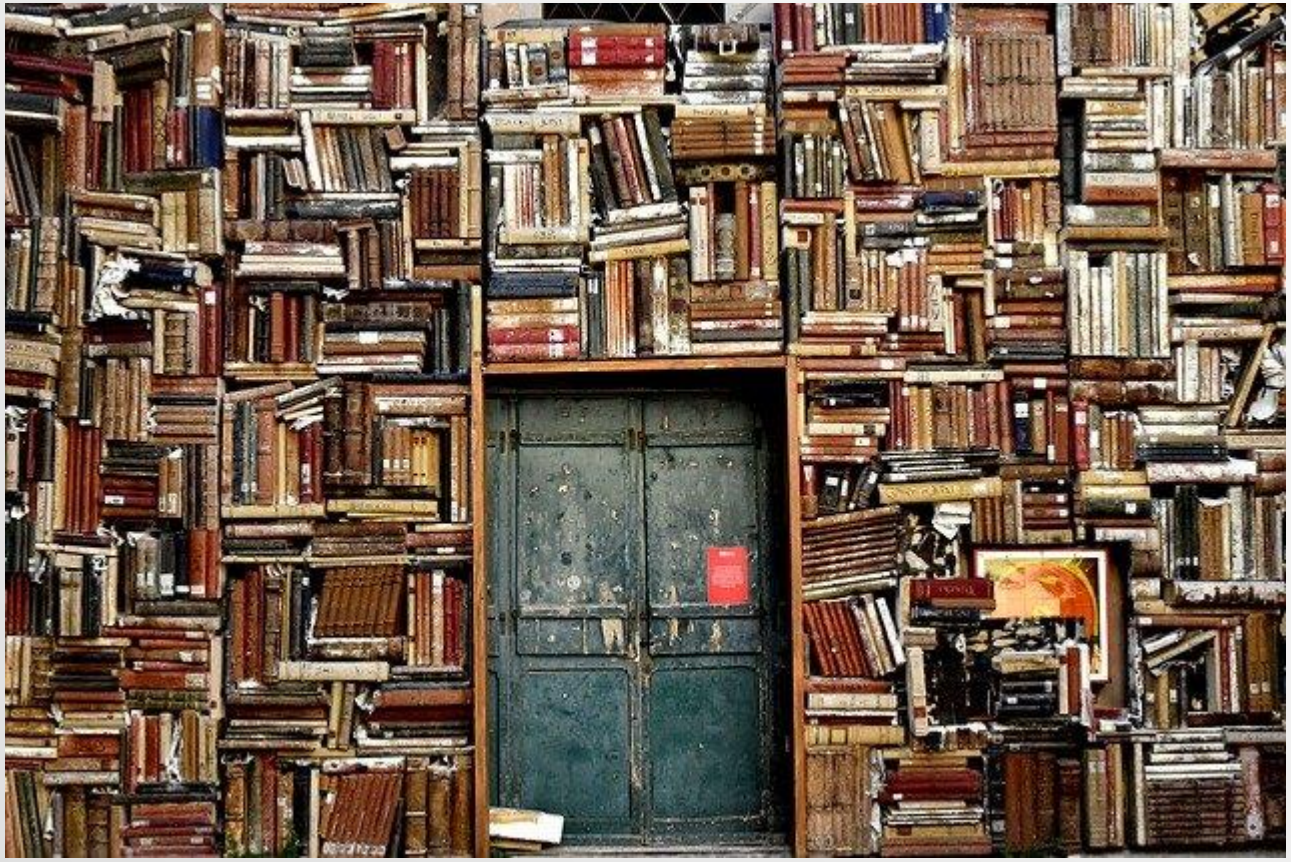
Transformers: antrenarea

- Antrenăm encoder pe un task general
  - Generăm o înțelegere a limbajului
  - Encoder-ul învață să transforme optim cuvintele în vectori
  - Encoderul învață să rezolve polisemantism, cuvinte subînțelese, metafore și contexte comune
- Rafinăm decoderul
  - Folosim modelul de limbaj deja format
  - Optimizăm pentru taskul dorit



Transformers: datele

- *Self-supervised learning:* nu necesită seturi de date adnotate
- Este nevoie de mult text
- Modele sunt antrenate pe Wikipedia, cărți întregi, articole, blog-uri etc.



Transformers – exemple de antrenare

- Avem o propoziție
  - Ascundem din ea 15% din cuvinte
  - Model trebuie să ghicească ce cuvânt ar trebui să fie acolo
  - Astfel modelul învață contextul cuvintelor
- Avem două propoziții
  - Modelul trebuie să ne spună dacă a doua are sens să urmeze după prima.
  - Astfel modelul învață contextul între propoziții
- Avem n propoziții,
  - Ascundem 15% din ele
  - Model trebuie să le ghicească în întregime!
  - Astfel modelul învață să fie 'creativ' și să poată abstractiza și genera propoziții întregi

Transformers: rafinare

- Encoderul antrenat reprezintă numeric cuvinte și propoziții
  - Ține cont și de context
- Rafinarea modelului pentru ...
  - Traducere
  - Rezumare
  - Generare de text
  - Predicția următoarelor n cuvinte
  - Răspunsuri la întrebări
  - Clasificarea documentelor
  - Recunoașterea entităților din propoziții (persoane, organizații, locații, etc)

Transformers: accesibilitate

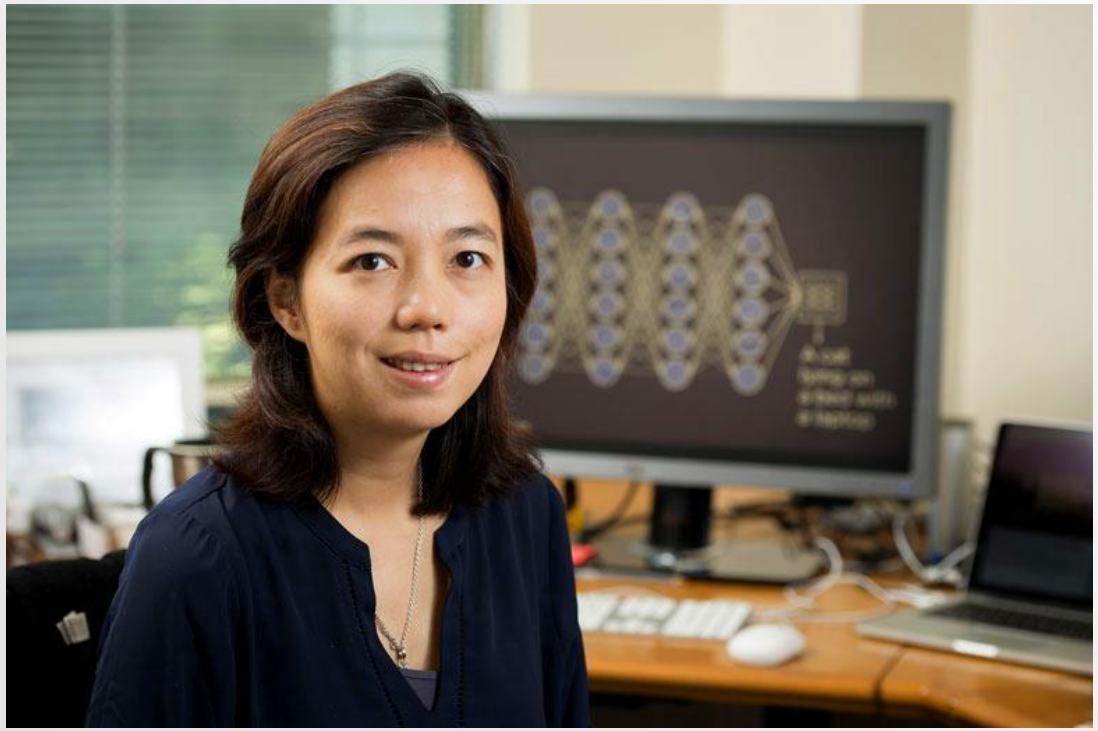
- Extrem de costisitoare de antrenat
  - Zeci pe zile pe sute de GPU-uri
- Dar: sunt modele empatice / reutilizabile
  - Trebuie antrenate o singură dată,
  - Pot fi transferate rapid către un task nou
- Actualmente accesibile pentru oricine
  - Descărcăm un model preantrenat pe orice limbă
  - Îl ajustăm în câteva ore pe procesorul personal
  - Rezolvăm o problemă nouă

#4.3 Procesarea imaginilor

Computer Vision: de la CNNs la Transformers?

Computer vision - origini

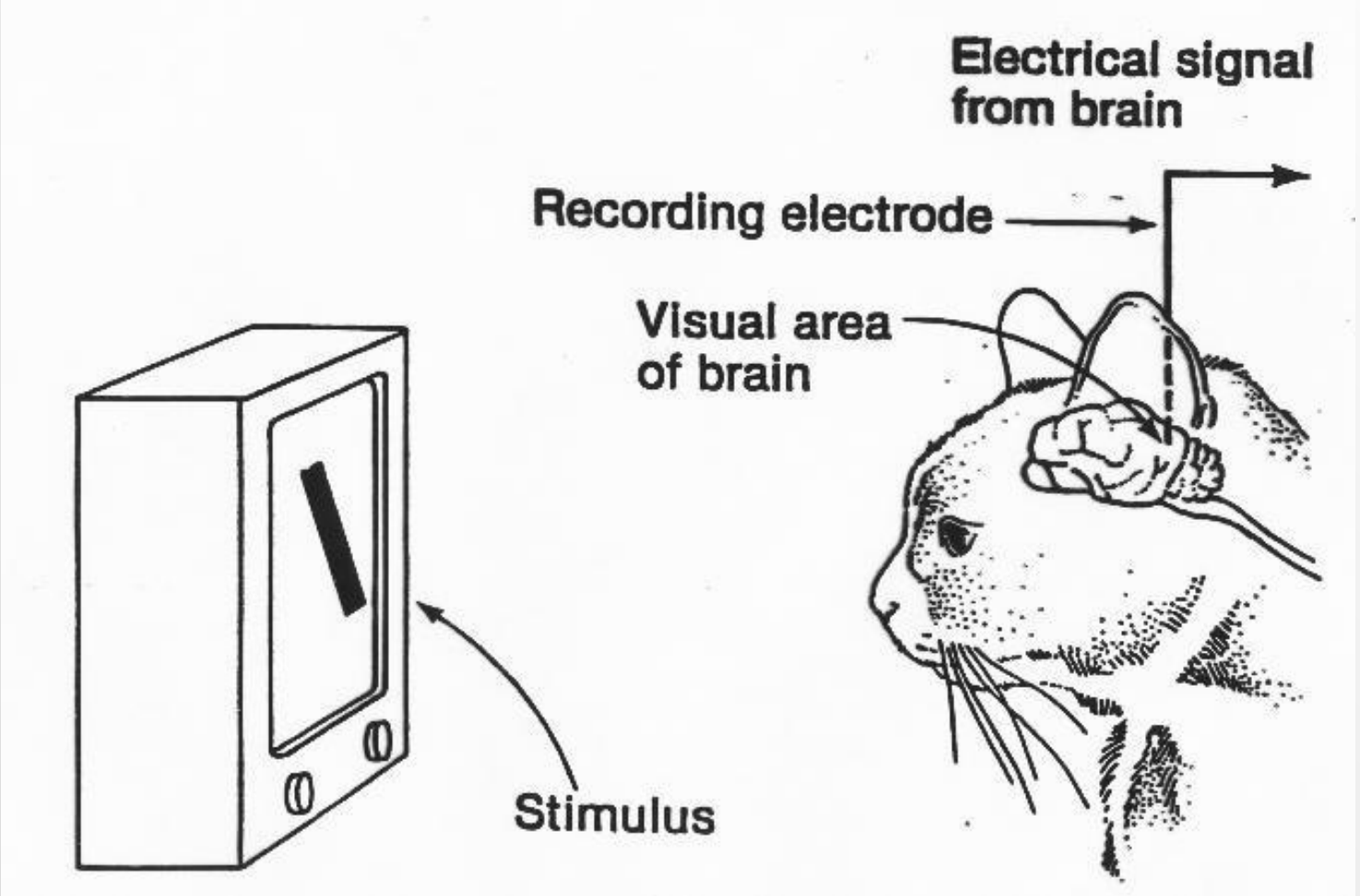
- A explodat odată cu ImageNet
  - Bază de date pentru recunoașterea imaginilor + concurs software anual: ImageNet Large Scale Visual Recognition Challenge (ILSVRC)
  - Fondat de Fei-Fei Li în 2012
  - Actualmente 14 mil.+ imagini adnotate
- Bazat pe o istorie lungă a procesării imaginii



Fei-Fei Li, Stanford University

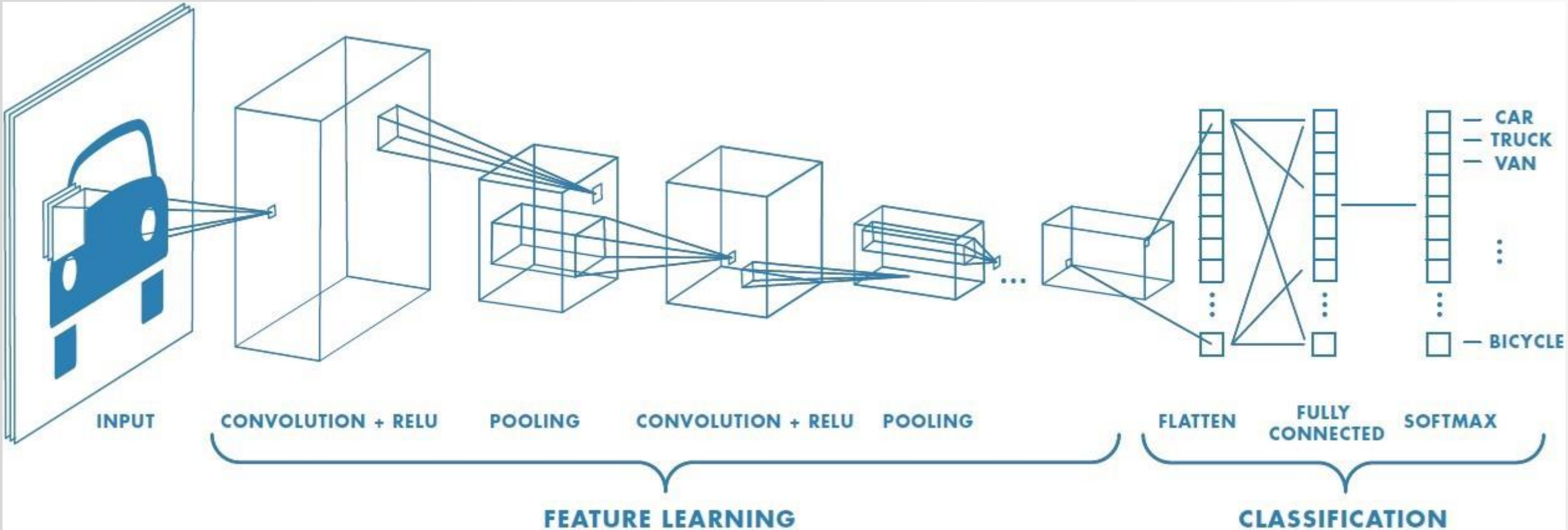
Computer vision – de la neuroni…

- *“Receptive fields of single neurons in the cat’s striate cortex”* (1959)
- Două tipuri de neuroni în cortexul vizual
  - Neuroni simpli: muchii și linii orientate, fixe în spațiu
  - Neuroni complecși: forme mutate în spațiu: invarianți spațial



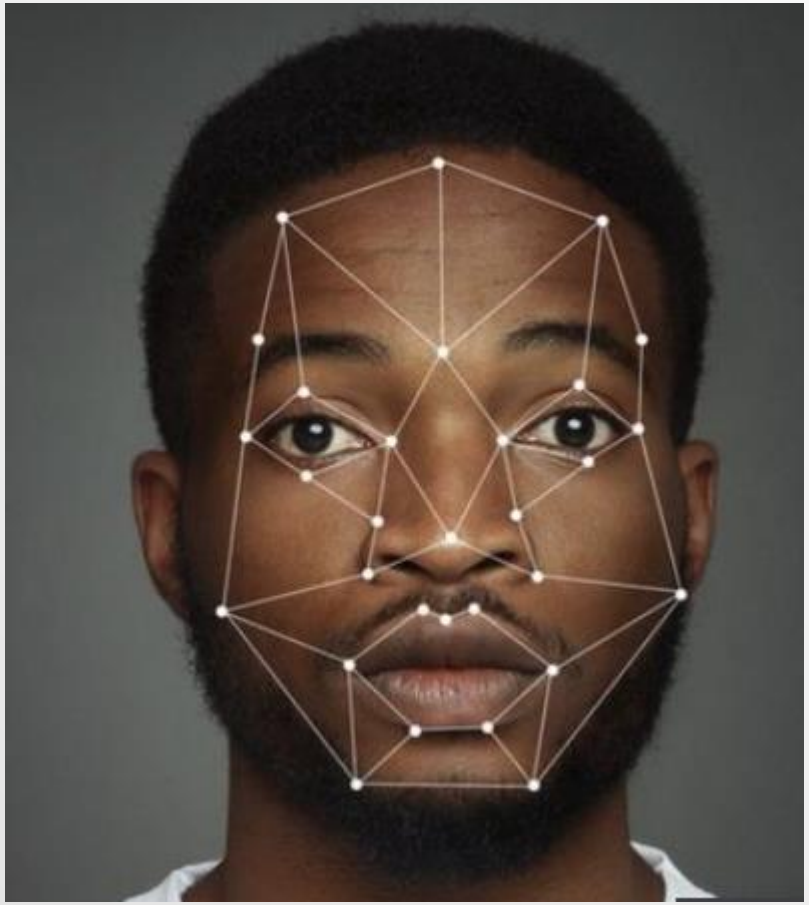
… la CNN-Rețele Neurale Convoluționale

- Stau la baza computer vision actual



CNN - Rețele Neurale Convoluționale

- Permit...
  - Identificare facială
  - Generare de fețe
  - Clasificare obiectelor din imagini
  - Completarea de părți lipsă din imagini
  - Scoaterea obstacolelor din imagini (de exemplu un gard)

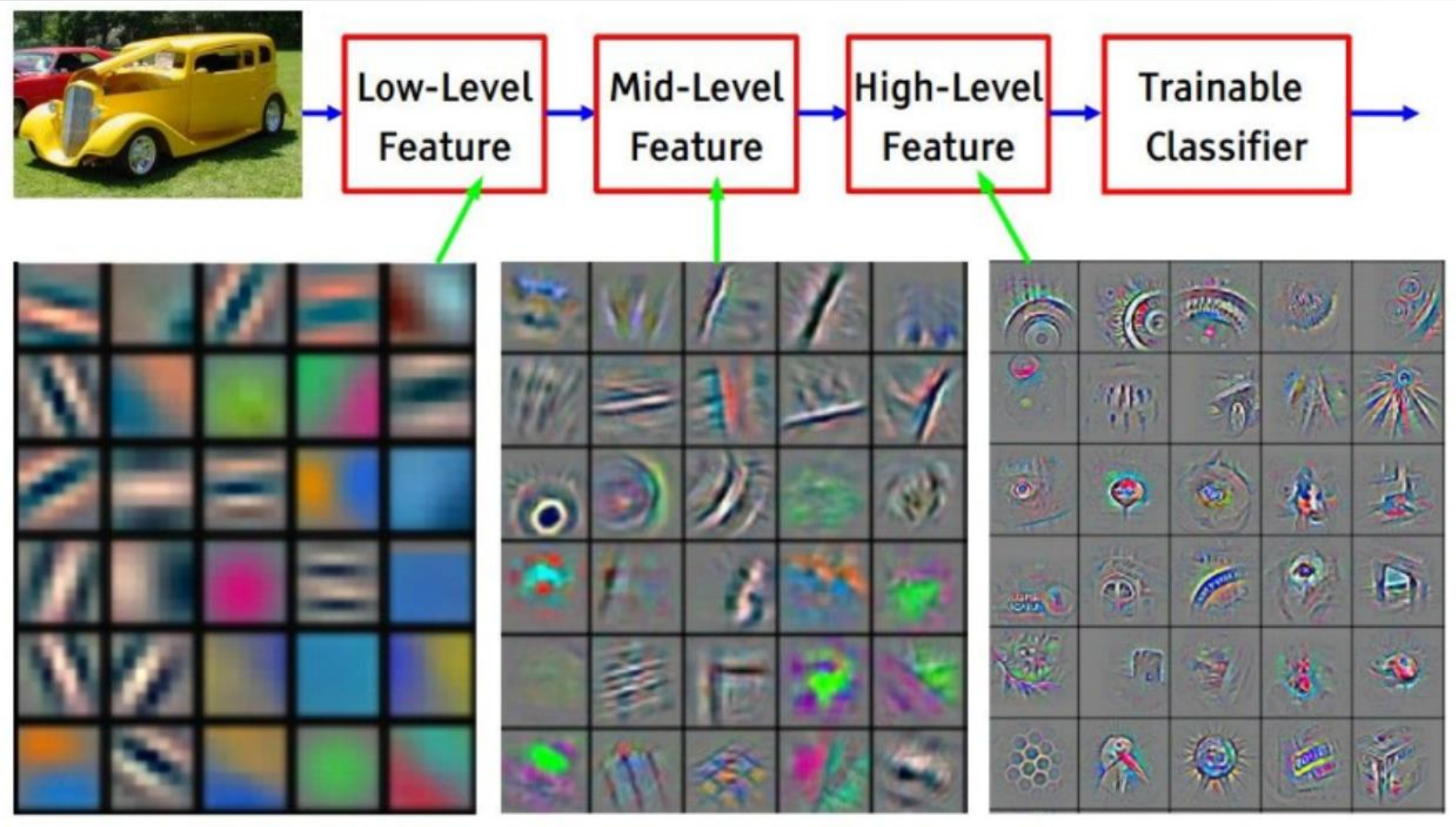


CNN – Funcționare

- Procesare secvențială, pe cadrane, a imaginii
- De la forme generale la detalii tot mai mici
  - Ce forme mari sunt în fiecare cadru?
  - Ce detalii? Tot mai fine

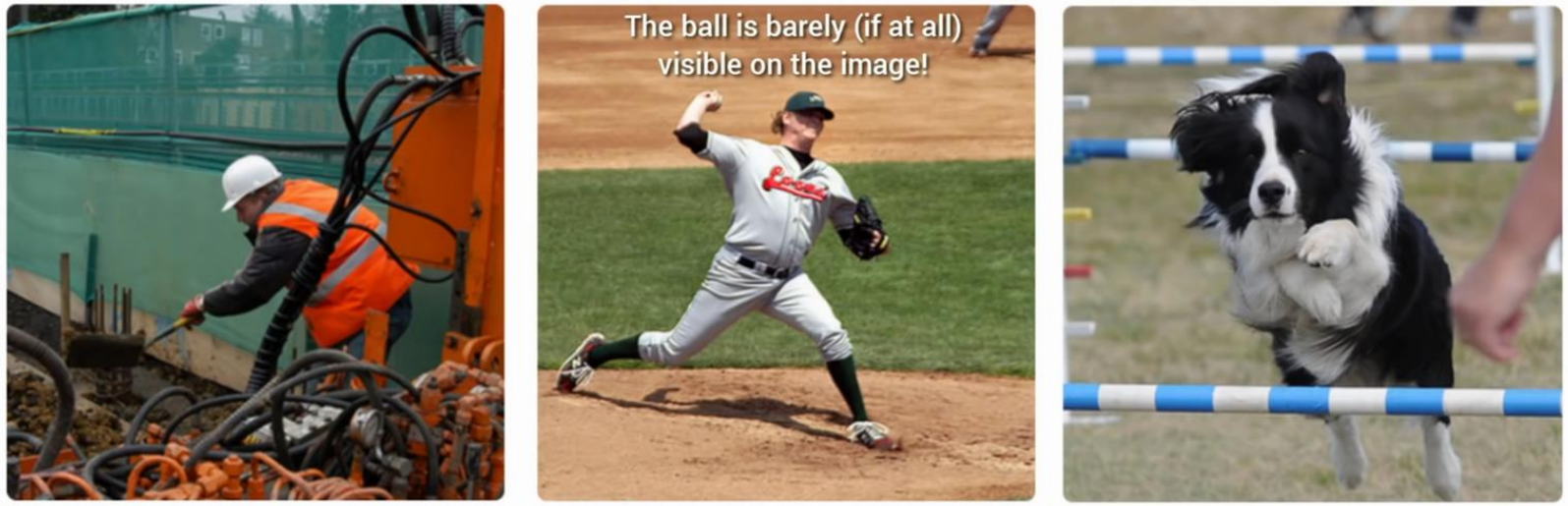
CNN – Detecția unei mașini

- De la general la specific
  - Modelul vede prima dată forma generală, precum parbrizul, capota, culoarea mașinii
  - Apoi se uită mai în detaliu, și observă roțile, farurile, geamurile
  - El poate intra și mai mult în detaliu, să se uite la oglinzi, mânere, etc
- Similar cu encoderul din transformere


Low-Level Feature
Mid-Level Feature
High-Level Feature
Trainable Classifier

CNN–Rezultate|1

- 2015: Identificarea și descrierea imaginilor (Karpathy & Fei-Fei)
- Probleme pentru obiecte multiple în cadru



"construction worker in orange safety vest is working on road."

"baseball player is throwing ball in game."

"black and white dog jumps over bar."

CNN–Rezultate|2

- 2017: Pornind de la o descriere, generăm o imagine complet nouă! (Reed et al.)
- Milestone important
- Limitări
  - Rezoluție mică
  - Schimbări mici de formulare au impact mare

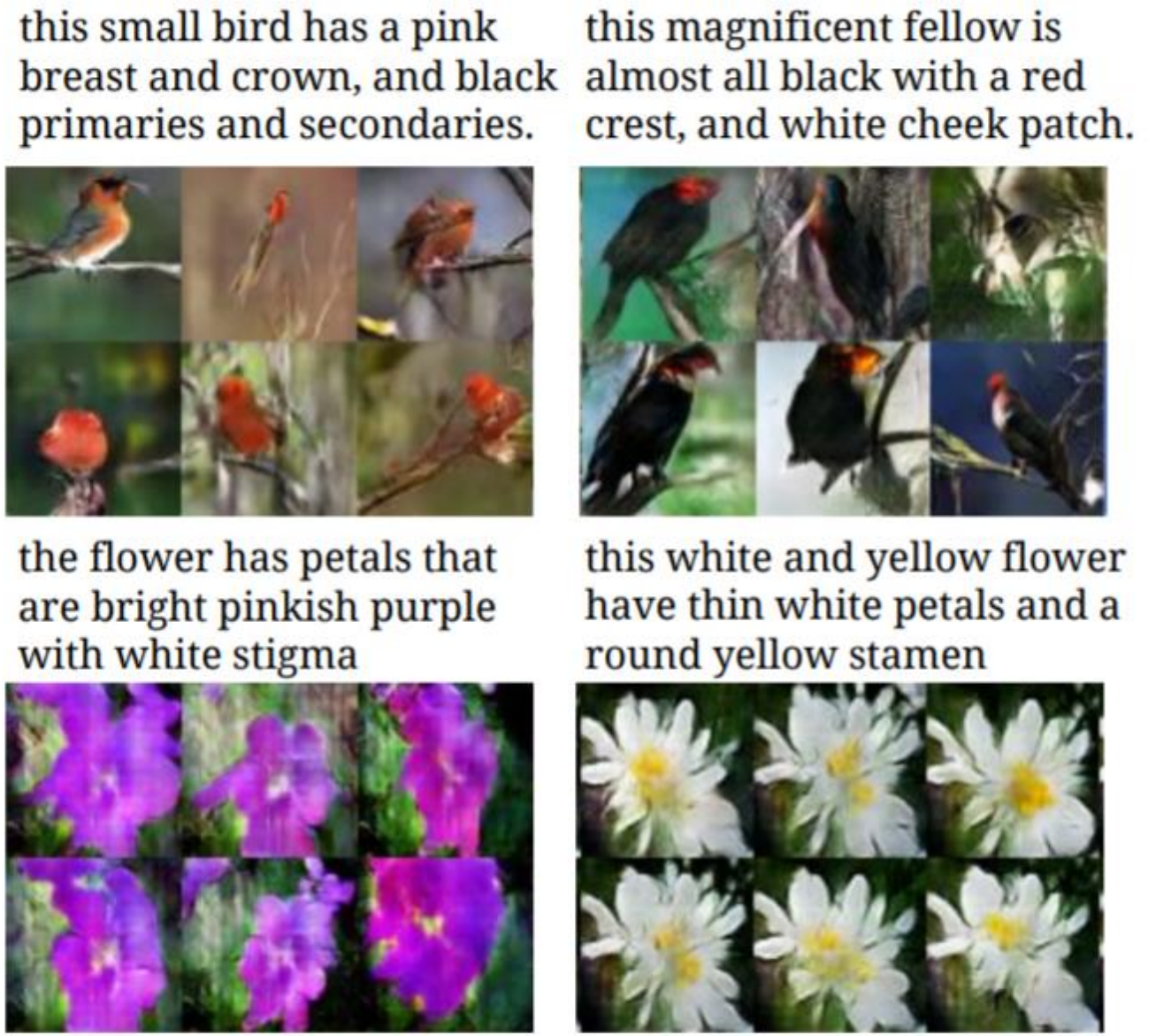
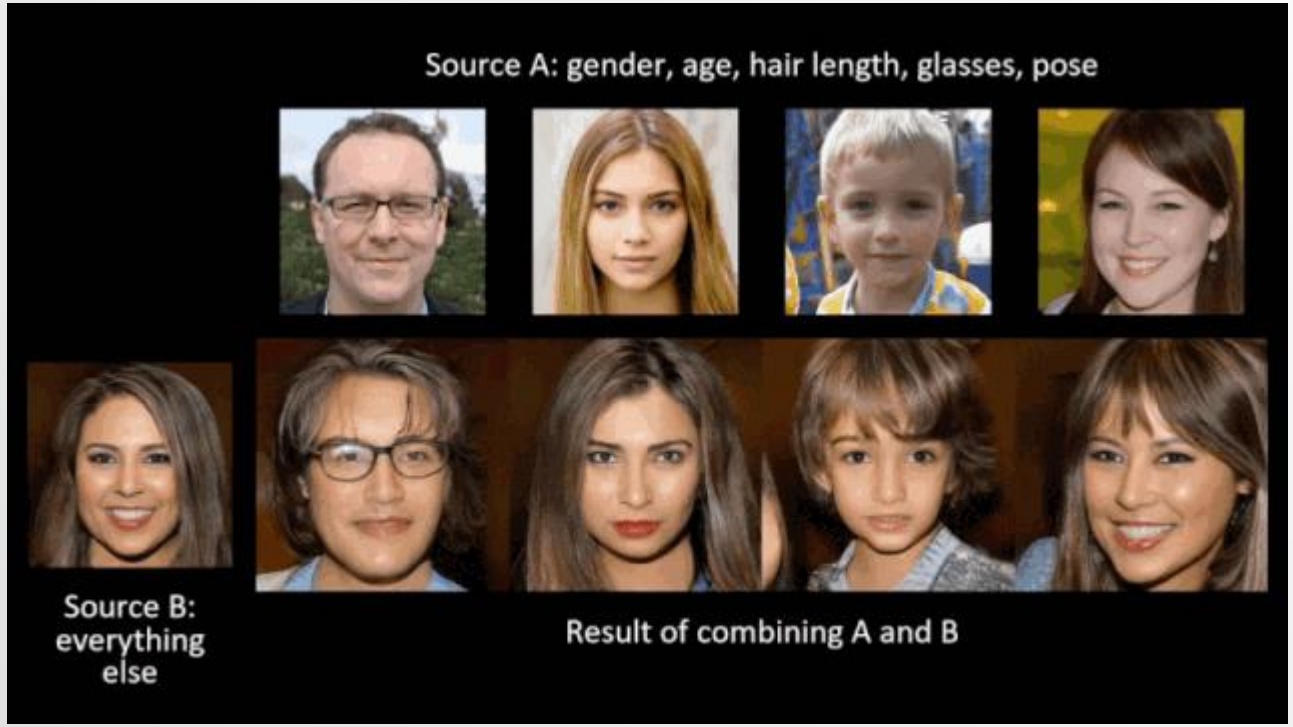


Figure 1. Examples of generated images from text description Left: captions are from zero-shot (held out) categories, unsee text. Right: captions are from the training set.

CNN-Rezultate|3

- 2018: StyleGAN – generarea de imagini bazate pe stiluri (Karras et al.)
  - Rezoluție mai bună
  - Control sporit asupra stilului imaginii
  - Limită: defecte în zone de detaliu (ochi, dinți)

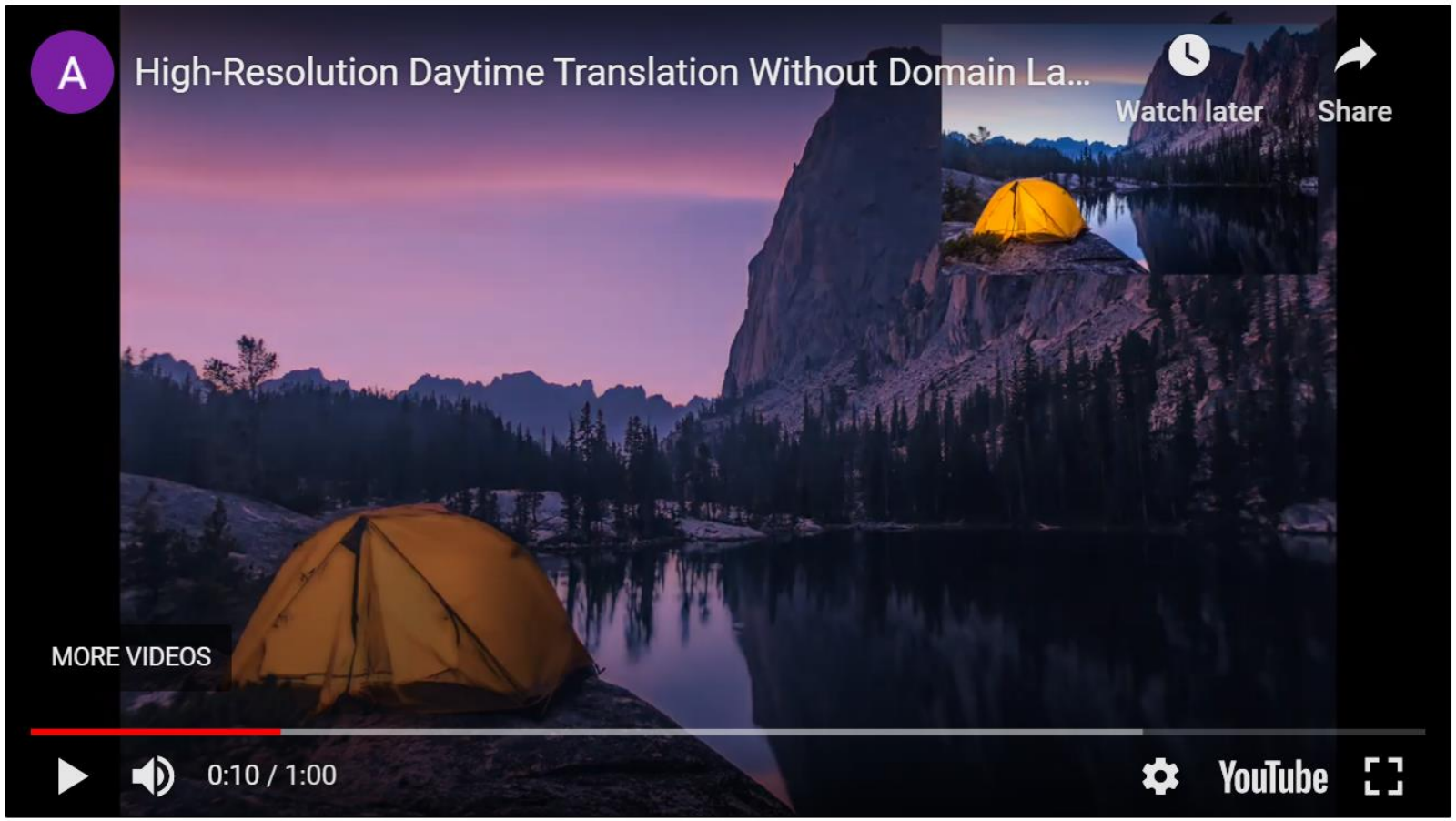


CNN-Rezultate|4

- 2019: StyleGAN 2 (Karras et al.)
- O creștere rapidă a calității în 2020, cu rezultate remarcabile

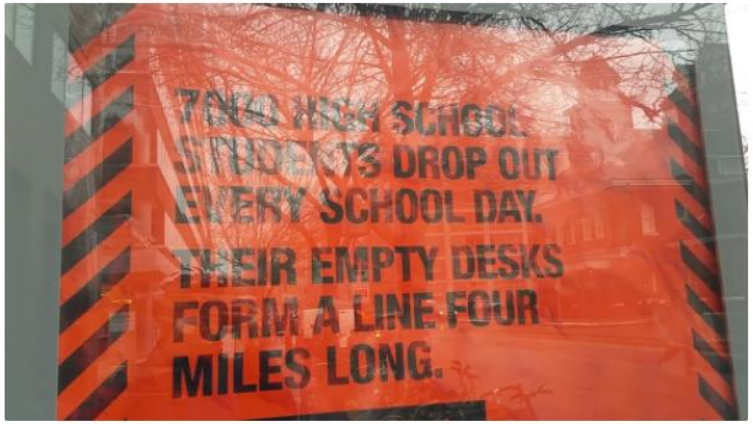
CNN-Rezultate|5

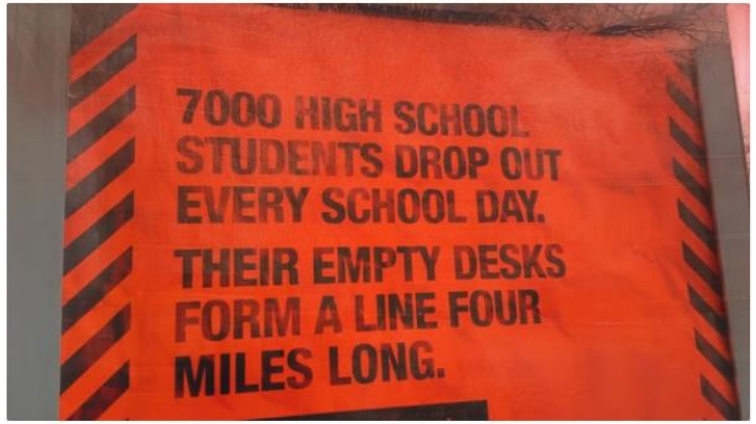
- Transformări de la zi la noapte
- Antrenarea a fost făcut fără imagini adnonatate!
- Stilurile sunt ușor transferabile
- Video:



CNN-Rezultate|6

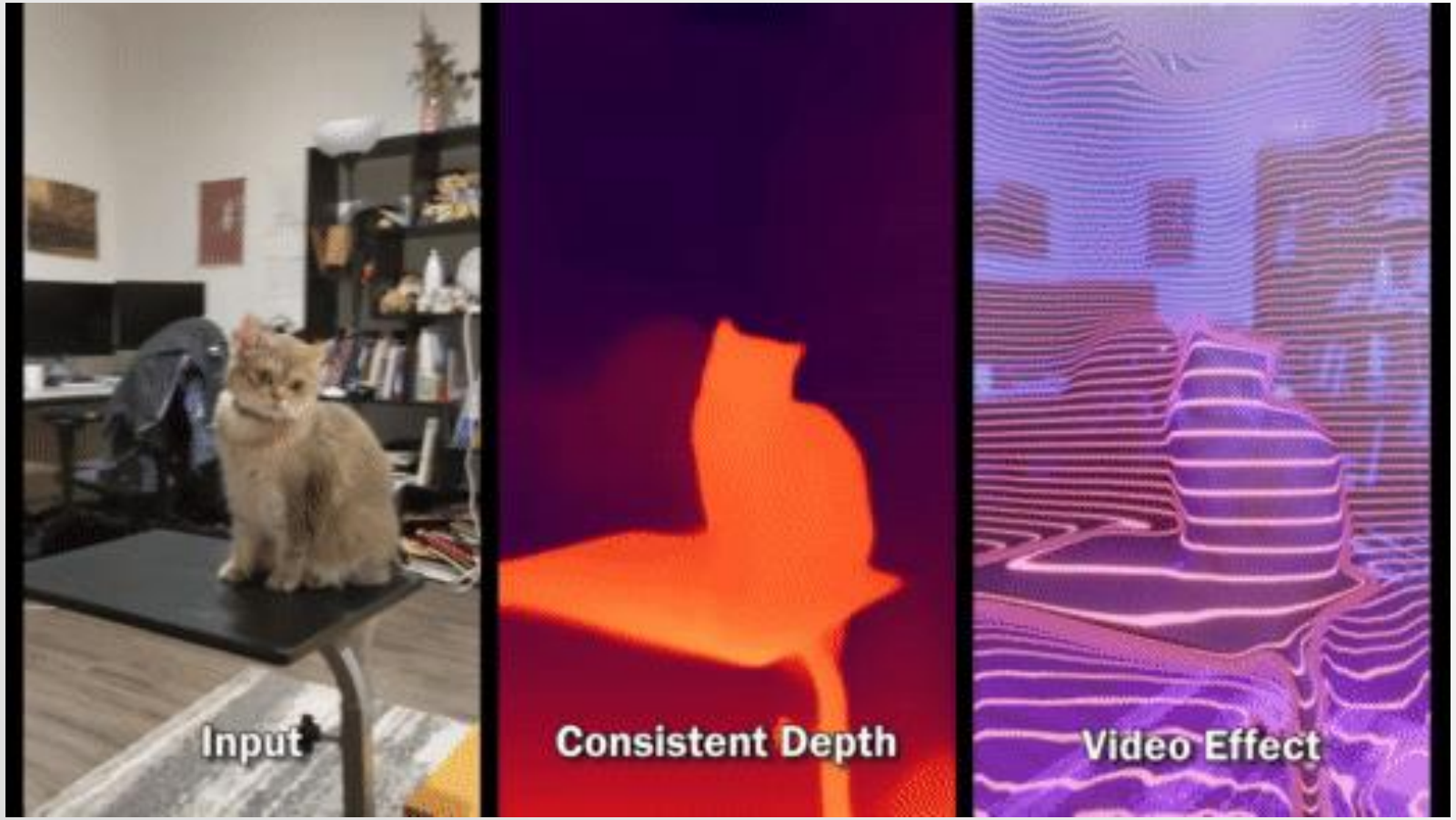
- Eliminarea obstacolelor (Liu et al.)
  - Gard
  - Reflexii
  - Picături de ploaie





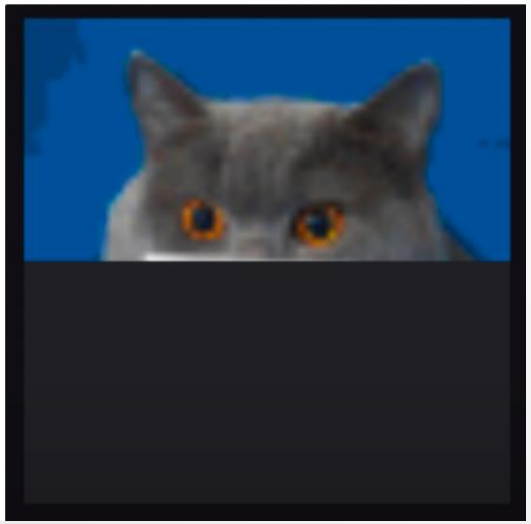
CNN-Rezultate|7

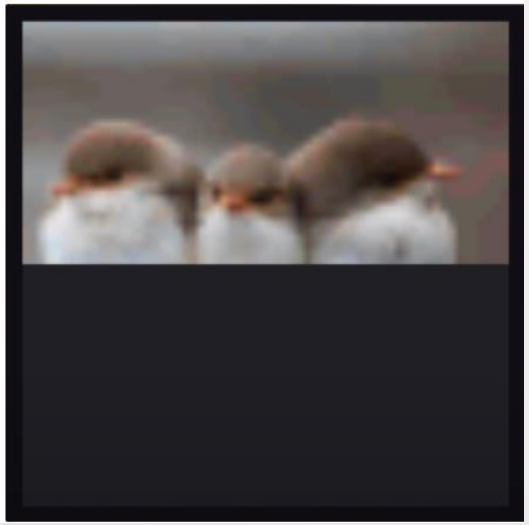
- Estimarea consistentă a adâncimii din video (Luo et al.)
  - Gard
  - Reflexii
  - Picături de ploaie



Ce urmează?

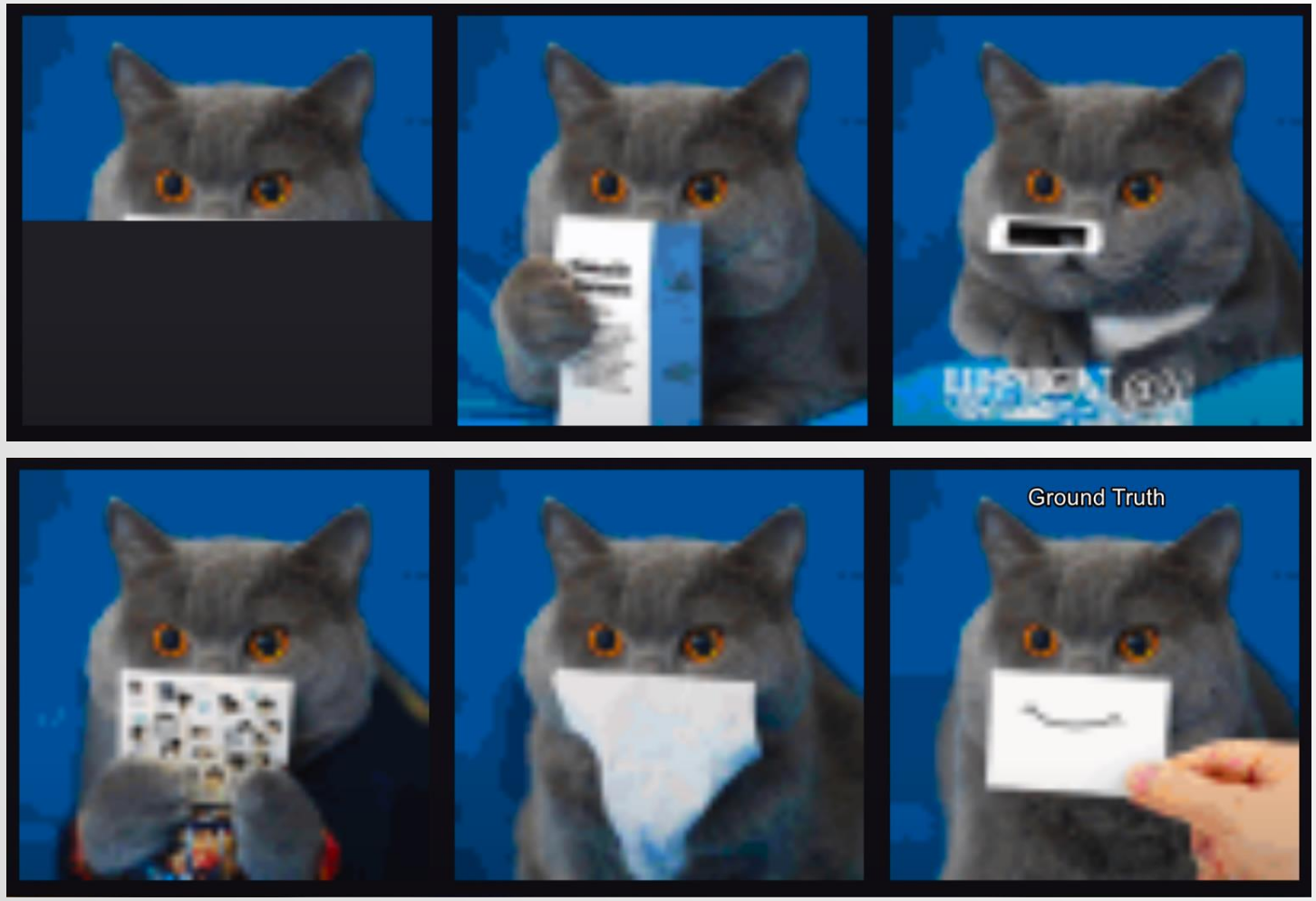
- Transformere pentru imagini?
- Transformerele au limite stricte de memorie
  - Fiecare input este legat cu toate celelalte în model: spațiul crește cu N^2, unde N este numarul de inputuri
- Aproape toate transformerele au o limită de input de 512
  - Pentru *cuvinte* este suficient
  - Pentru pixelii unei imagini e prea puțin
- Totuși...

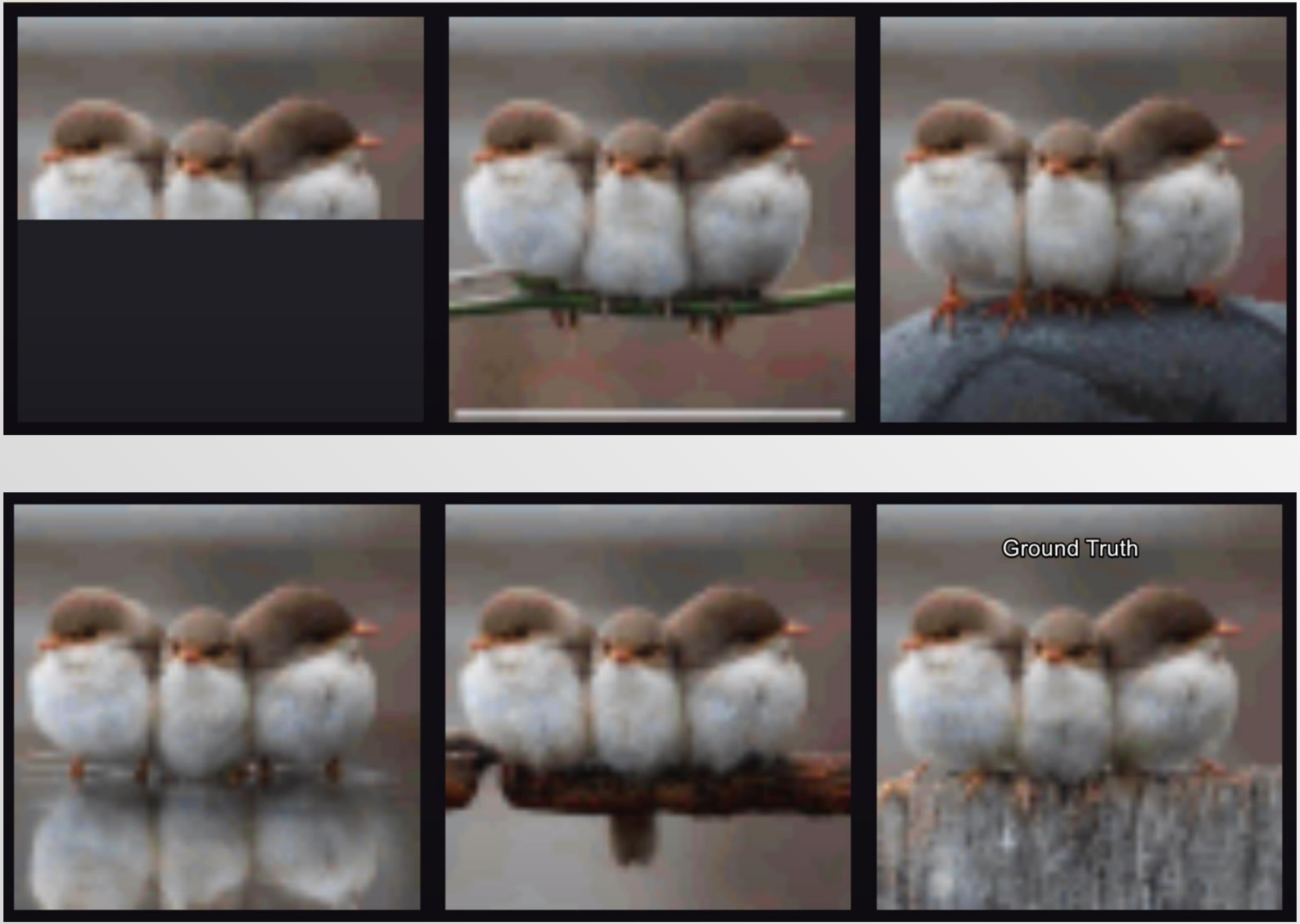




Transformers: ImageGPT

- Completarea imaginii cu ImageGPT (Chen et.al)
- Modelul primește o imagine incompletă, și i se cere să o umple cu informație relevantă




Ground Truth

Transformers – Evoluții actuale

- **Vor fi CNN-urile înlocuite de transformere odată ce avem mai multă memorie disponibilă?**
- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Oct. 2020)
  - O abordare pură de transformere
  - Fără nevoia unui model gigantic precum GPT-3 pentru imagini
  - Optimizare modulului de atenție

Pe scurt: ML și Deep Learning

4.1 Procesarea vocii

- Audrey: problemele predicției modulare
- Rețele neurale: avantaje și probleme

4.2 Procesarea Limbajului Natural (NLP)

- De la Bag of Words la Transformers

4.3 Procesarea Imaginilor

- Convoluted Neural Networks (CNNs)
- Transformers pentru imagini?

Referințe

- Lynskey, D. 2019. 'Alexa, are you invading my privacy?' – the dark side of our voice assistants. The Guardian
- George, S. 2020. A Natural Transition from “Bag of words” to “Transformers” — A Recap of NLP’s Journey So Far. Medium
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
- Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396.
- Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4401-4410).

Referințe | 2

- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8110-8119).
- Liu, Y. L., Lai, W. S., Yang, M. H., Chuang, Y. Y., & Huang, J. B. (2020). Learning to See Through Obstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14215-14224).
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396.
- Luo, X., Huang, J. B., Szeliski, R., Matzen, K., & Kopf, J. (2020). Consistent video depth estimation. arXiv preprint arXiv:2004.15021.
- Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Dhariwal, P., Luan, D. and Sutskever, I., 2020, November. Generative pretraining from pixels. In Proceedings of the 37th International Conference on Machine Learning
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

SYSTEMS LABORATORY

Cursul DS05

Predicția

PR network crunch

Structura cursului

- Introducere: De ce DS?
- Scopul cursului
- Măsurarea și colectarea datelor
- Erori de măsurare
  - Precizia
  - Validitatea
  - Fidelitatea
- Corelația
- Erori de modelare
  - Corelații iluzorii
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erori de eșantionare
- Tehnologiile datelor: programare vs. Învățare automată
- Procesul învățării automate
- Machine Learning și Deep Learning
- Predicție prin analiza datelor
  - Analiza factorială (reducerea dimensionalității)
  - Clasificare (cluster)
  - Analiza de rețea
  - Analiza de regresie
  - Predicții și serii de timp
- Predicția în capitalismul datelor

Introducere

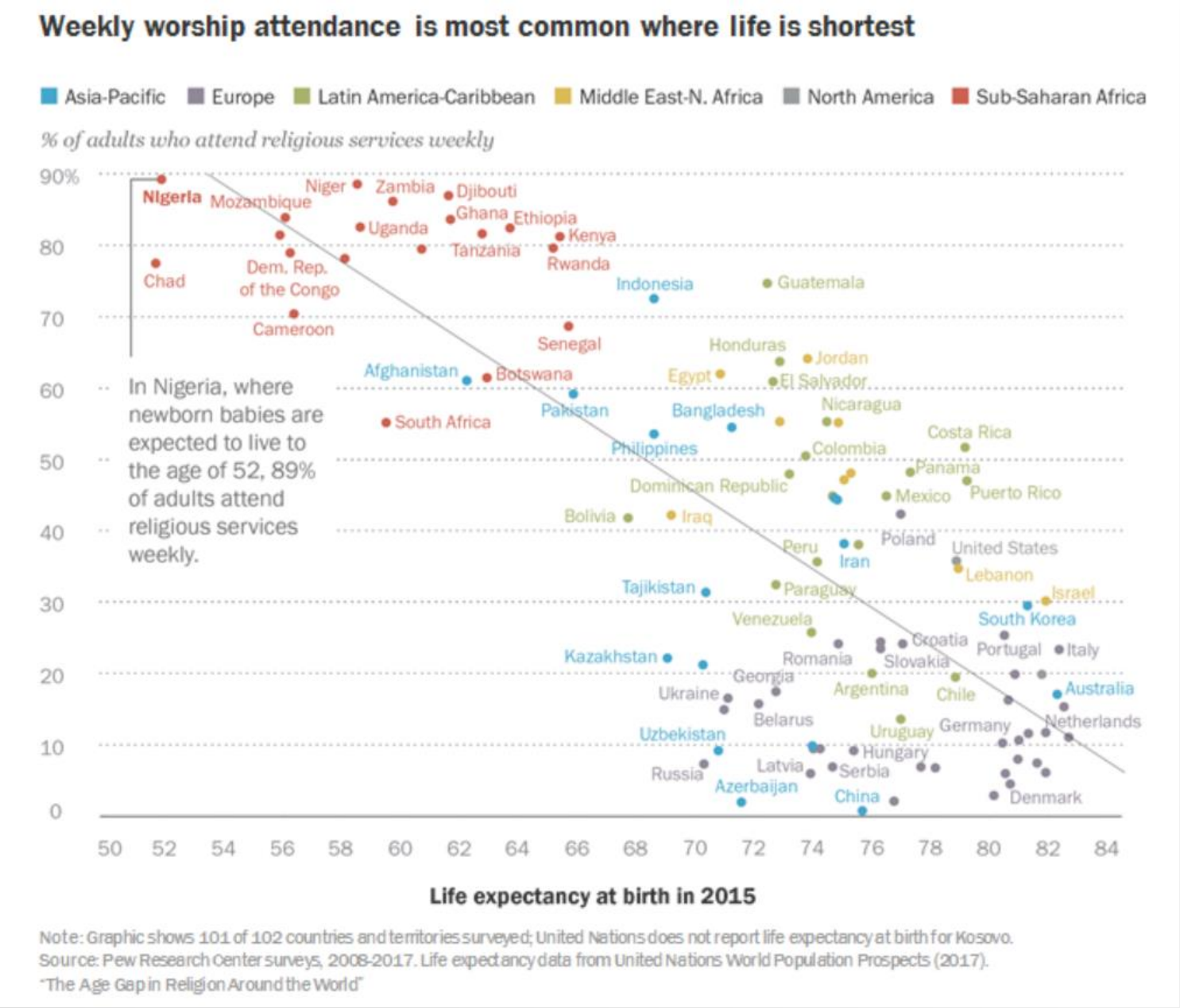
- Predicția / anticiparea este necesară acțiunii
- Tehnicile de analiză a datelor pot identifica patternuri
  - Modelele „prezic” datele, adică reproduc patternuri trecute
  - Oamenii prezic / anticipează viitorul, pornind de la modele și asumpții politice despre viitor
- Tehnici de analiză a datelor:
  - Reducerea dimensionalității, analiză cluster, analiză de rețea
  - Analiză de regresie, serii de timp
- Controverse curente
  - Decizii algoritmice și părtinire algoritmică (algorithmic bias)
  - Capitalismul datelor, AdTech și autonomia individuală

#3.1 Analiza datelor

Predicție și cunoaștere

Model vs. predicție

- Modelul „prezice” datele
  - Modelul „prezice” trecutul
  - **Re-producerea** patternurilor trecute
- Noi prezicem viitorul folosind:
  - Modelul
  - Asumpții privind continuitatea trecut – prezent – viitor



Predicție vs. „algorithmic bias”

- „Părtinire algoritmică” (en. algorithmic bias)
  - Decizii bazate pe ML și Big Data
  - Modelele reproduc patternuri existente în date
  - Pattern = diferențe = inegalități = stereotipuri
- Exemple de părtinire algoritmică
  - UK, 2020: „prezicerea” notei la Bac pe baza prestigiului școlii
  - SUA, 2016: „prezicerea” probabilității de recidivă pe baza profilului social (implicit rasial) al indivizilor

MIT Technology Review

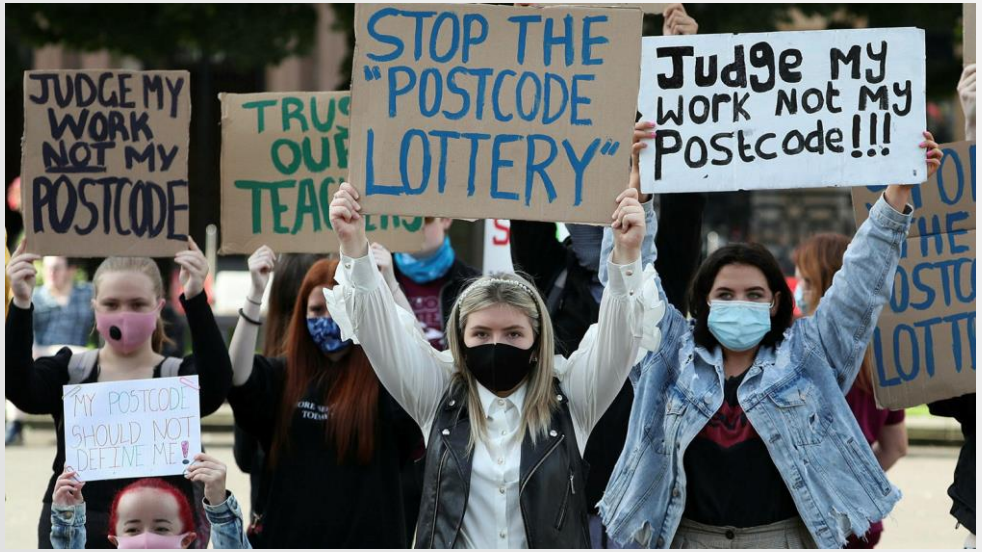
Tech policy / AI Ethics

The UK exam debacle reminds us that algorithms can't fix broken systems

The problem began when the exam regulator lost sight of the ultimate goal—and pushed for standardization above all else.

by **Karen Hao** August 20, 2020

(MIT)



- Approximately 39% of A-level results were downgraded by UK algorithm.
- Disadvantaged students were the worst affected as the algorithm copied the inequalities that exist in the U.K.'s education system.

(CNBC, Aug. 2020)

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ProPublica, 2016

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

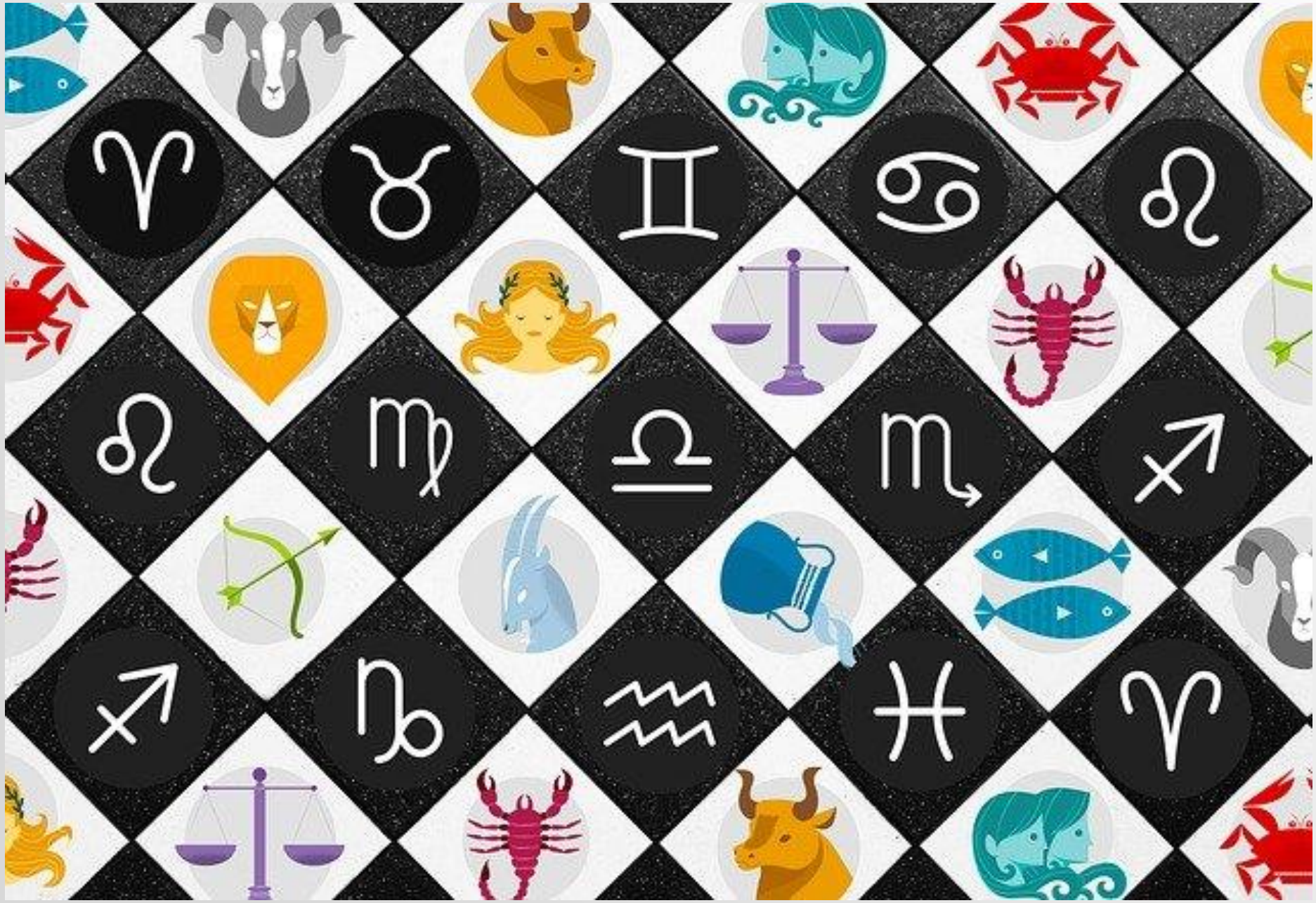
Model: predicție vs. descriere

ML & Analiza datelor

- Predicție = gestionarea complexității prezente și trecute
- „Modelul prezice datele” = „Modelul descrie datele”
- ML: Cât de bine prezice (ghicește, nimerește) modelul datele de test?
- Analiza datelor: Cât de bine prezice (ghicește, nimerește) modelul datele output ca funcție a datelor input?

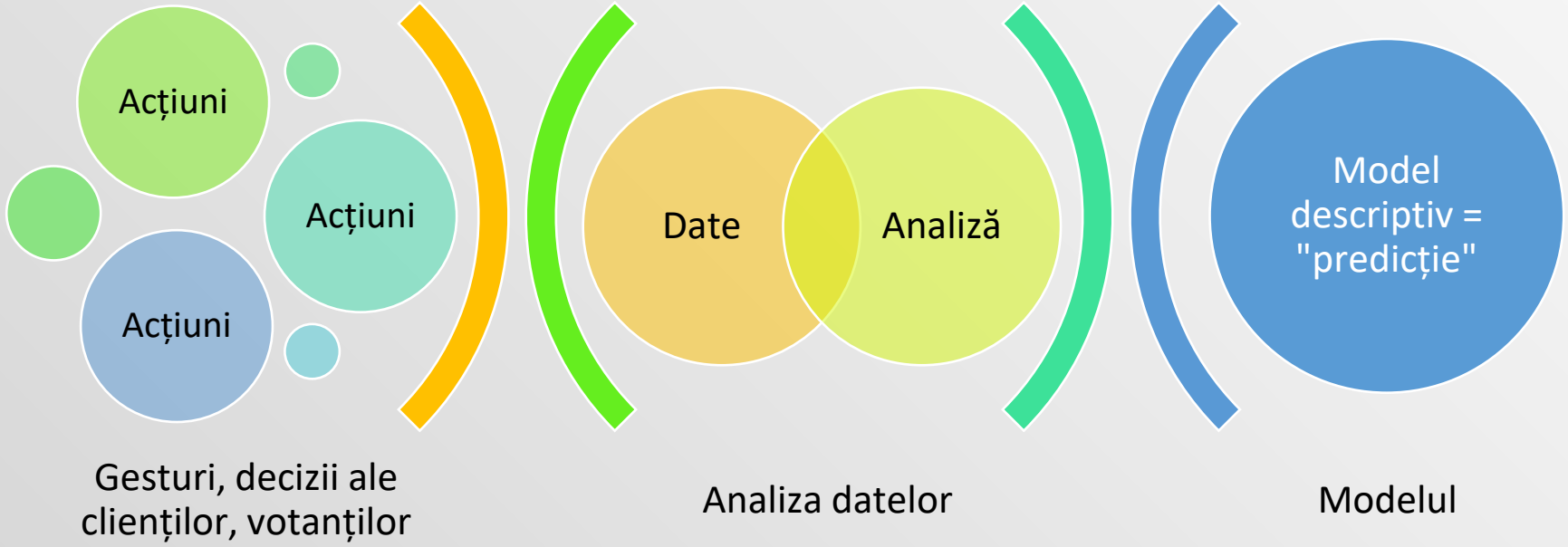
Viață

- Predicție = anticiparea viitorului
- Predicția se poate dovedi (in)corectă
- Predicția (co)produce realitatea
  - Acționăm pe baza unui model
  - Predicția auto-realizatoare / auto-negatoare



Ne determină zodiile personalitatea?

- Depinde dacă noi credem că determină!
- Pentru copii și cei ce nu sunt interesați de zodii, data nașterii nu corelează cu testele de personalitate
- Cu cât crezi mai mult în zodiac, corelează mai mult!
  - Studiile lui Hans Eysenck discutate în Richard Wiseman, Quirkology: How we discover the big truths in small things. Basic Books, 2007


Acțiuni
Acțiuni
Acțiuni
Date
Analiză
Model descriptiv = "predicție"
Gesturi, decizii ale clienților, votanților
Analiza datelor
Modelul

Acțiuni – date – analiză - model

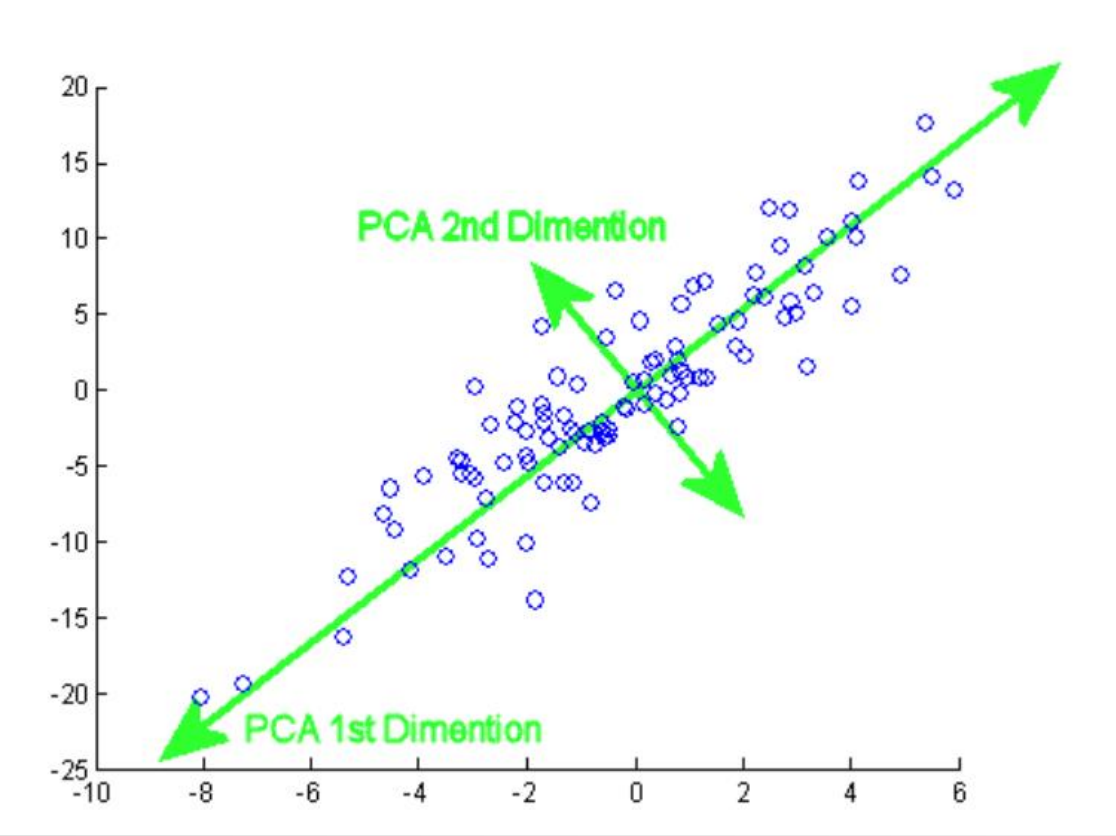
- „Predicție” = model retrospectiv pe acțiunile celorlalți
  - Ce fel de om ești? Ce fel de client ești? Ce fel de votant ești?
  - Cu ce anume corelează acțiunile trecute?
- Tehnici de analiză a datelor
  - Analiza factorială / Reducerea dimensionalității
    - Diagnoza personalității pe baza acțiunilor, preferințelor trecute
  - Analiza cluster
    - Diagnoza personalității pe baza celor care îți seamănă: „Spune-mi cu cin’ te-nsoțești, ca să-ți spun cine ești”
  - Analiza de rețea
    - Diagnoza personalității pe baza rețelei de apropiați: „Cin’ se-aseamănă, se-adună”
  - Analiza de regresie și seriile de timp
    - Explicarea acțiunilor prin factori externi și tendințe temporale

3.1.1 Analiza factorială

Reducerea dimensionalității

Reducerea dimensionalității

- Care sunt cele mai importate dimensiuni ale datelor mele?
  - Compresia datelor
  - Descoperirea structurilor fundamentale
  - Eliminarea erorilor de măsurare
  - Vizualizarea volumelor uriașe de date
  - Extragerea de date pentru antrenarea altor modele



Sursă

Big Five și măsurarea personalității

Low	Traits	High
Down to earth, pragmatic, risk averse, rational, conservative, straightforward	**Openness** Measures how creative, imaginative, down to earth or pragmatic someone is.	Creative, curious, sensitive to aesthetics, receptive to change, tolerant, liberal
Content, bold, carefree, easygoing, spontaneous, creative	**Conscientiousness** Conscientiousness measures preference for an organized approach to life in contrast to a spontaneous one.	Organised, reliable, consistent, enjoy planning, seek achievement
Solitary, reserved, relaxed, serious, team-orientated	**Extraversion** Extraversion measures a tendency to seek stimulation in the external world, the company of others, and to express positive emotions.	Outgoing, sociable, friendly, talkative, energetic, inclined to leadership
Assertive, skeptical, devious, autonomous, indifferent, uncompromising	**Agreeableness** Agreeableness relates to a focus on maintaining positive social relations, being friendly, compassionate, and cooperative.	Friendly, compassionate, gullible, cooperative, trusting, ready to compromise
Calm, self-confident, bold, emotionally stable, carefree	**Neuroticism** Neuroticism measures the tendency to experience mood swings and emotions such as guilt, anger, anxiety, and depression.	Anxious, pensive, impulsive, self-conscious, yielding

citizen me

Source

Loneliness and the Big Five Personality Traits: A Meta-analysis

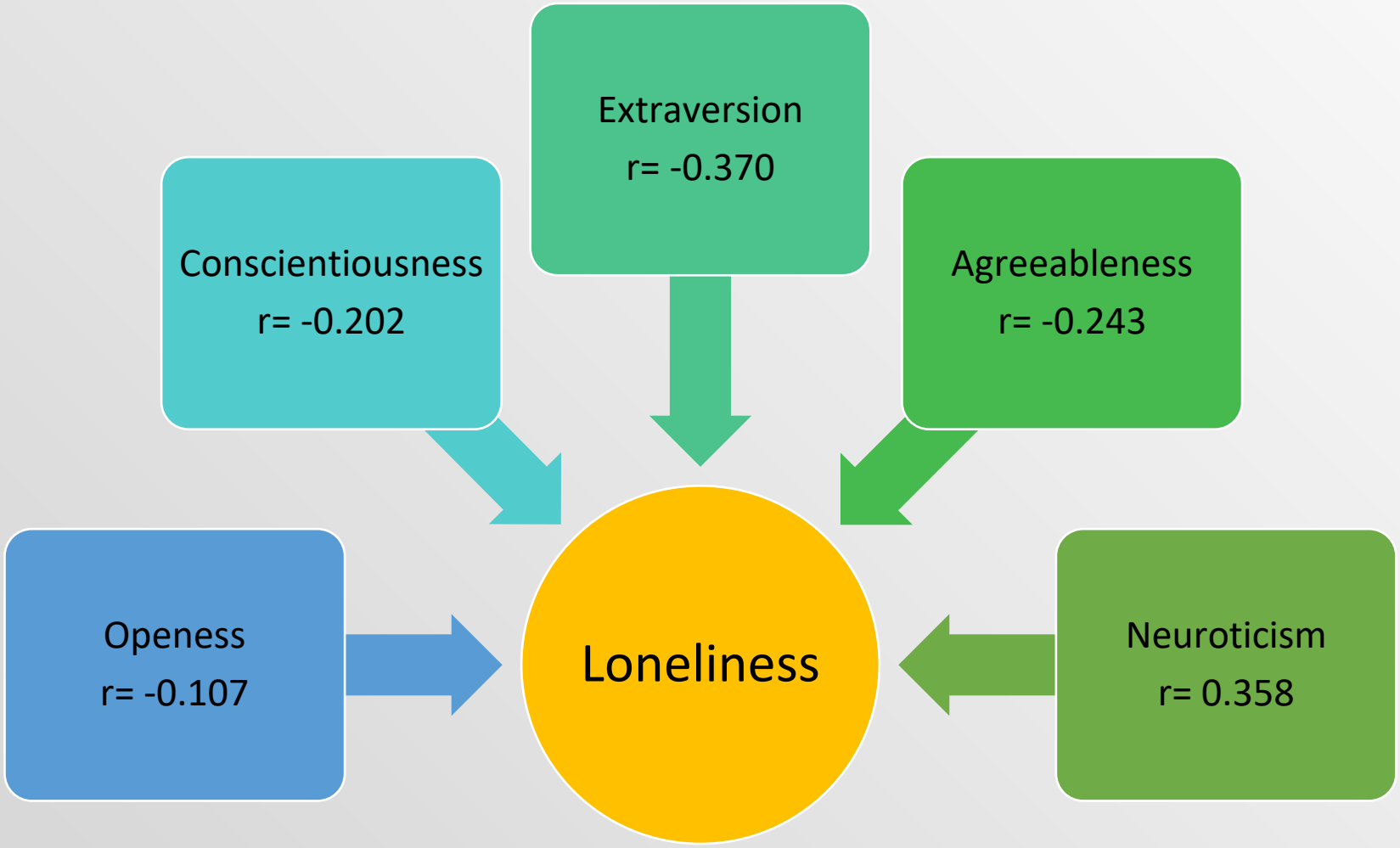
SUSANNE BUECKER[1]*, MARLIES MAES[2], JAAP J. A. DENISSEN[3] and MAIKE LUHMANN[1]

[1]Ruhr-Universität Bochum, Germany
[2]KU, Leuven, Belgium
[3]Tilburg University, The Netherlands

Abstract: This preregistered meta-analysis (k = 113, total n = 93 668) addressed how the Big Five dimensions of personality (extraversion, agreeableness, conscientiousness, neuroticism, and openness) are related to loneliness. Robust variance estimation accounting for the dependency of effect sizes was used to compute meta-analytic bivariate correlations between loneliness and personality. Extraversion (r = −.370), agreeableness (r = −.243), conscientiousness (r = −.202), and openness (r = −.107) were negatively related to loneliness. Neuroticism (r = .358) was positively related to loneliness. These associations differed meaningfully in strength depending on how loneliness was assessed. Additionally, meta-analytic structural equation modelling was used to investigate the unique association between each personality trait and loneliness while controlling for the other four personality traits. All personality traits except openness remained statistically significantly associated with loneliness when controlling for the other personality traits. Our results show the importance of stable personality factors in explaining individual differences in loneliness.
© 2020 European Association of Personality Psychology

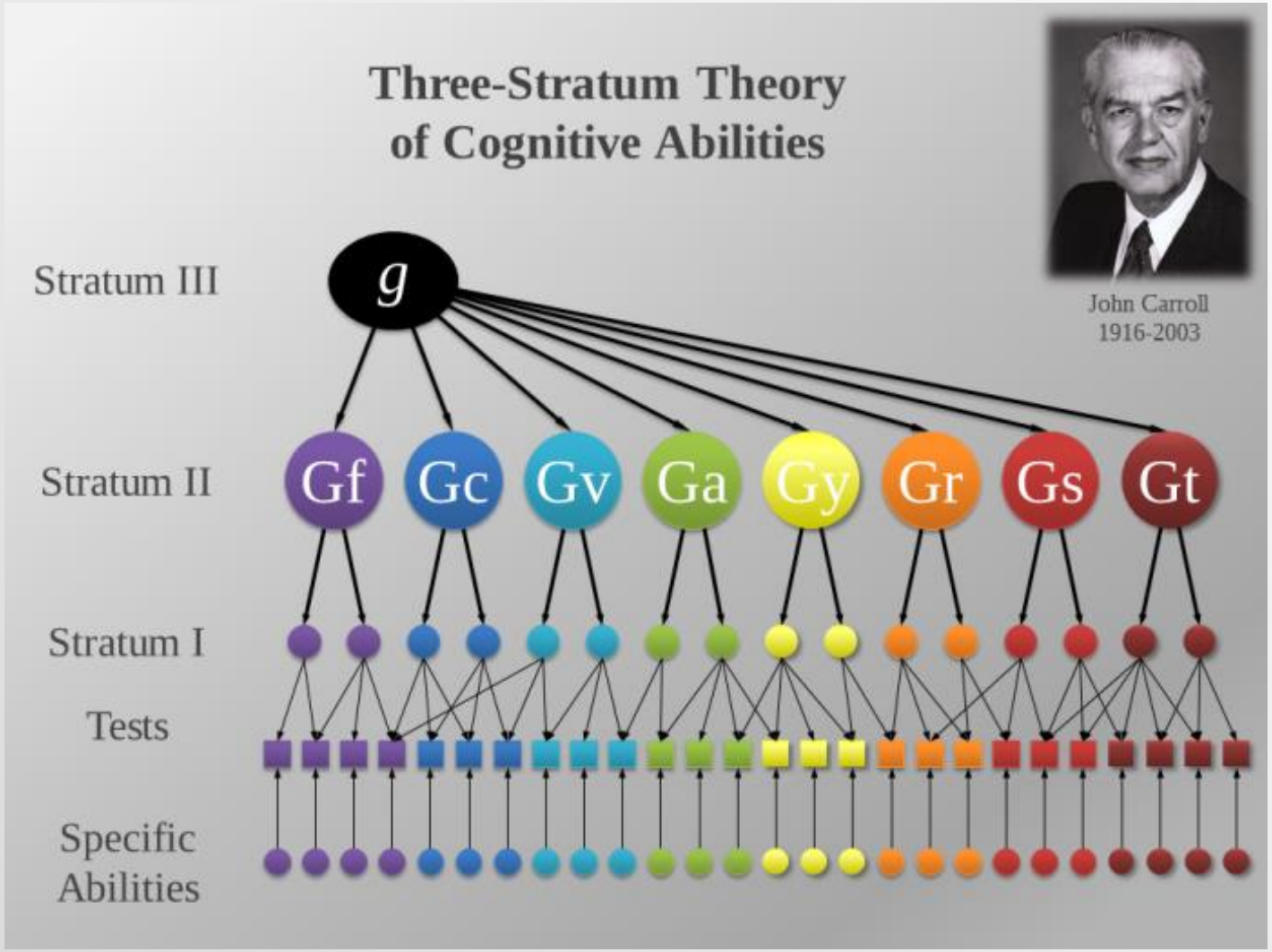
Key words: loneliness; perceived social isolation; personality; Big Five; meta-analysis

Buecker et al., 2020


Extraversion
r= -0.370
Conscientiousness
r= -0.202
Agreeableness
r= -0.243
Openess
r= -0.107
Loneliness
Neuroticism
r= 0.358

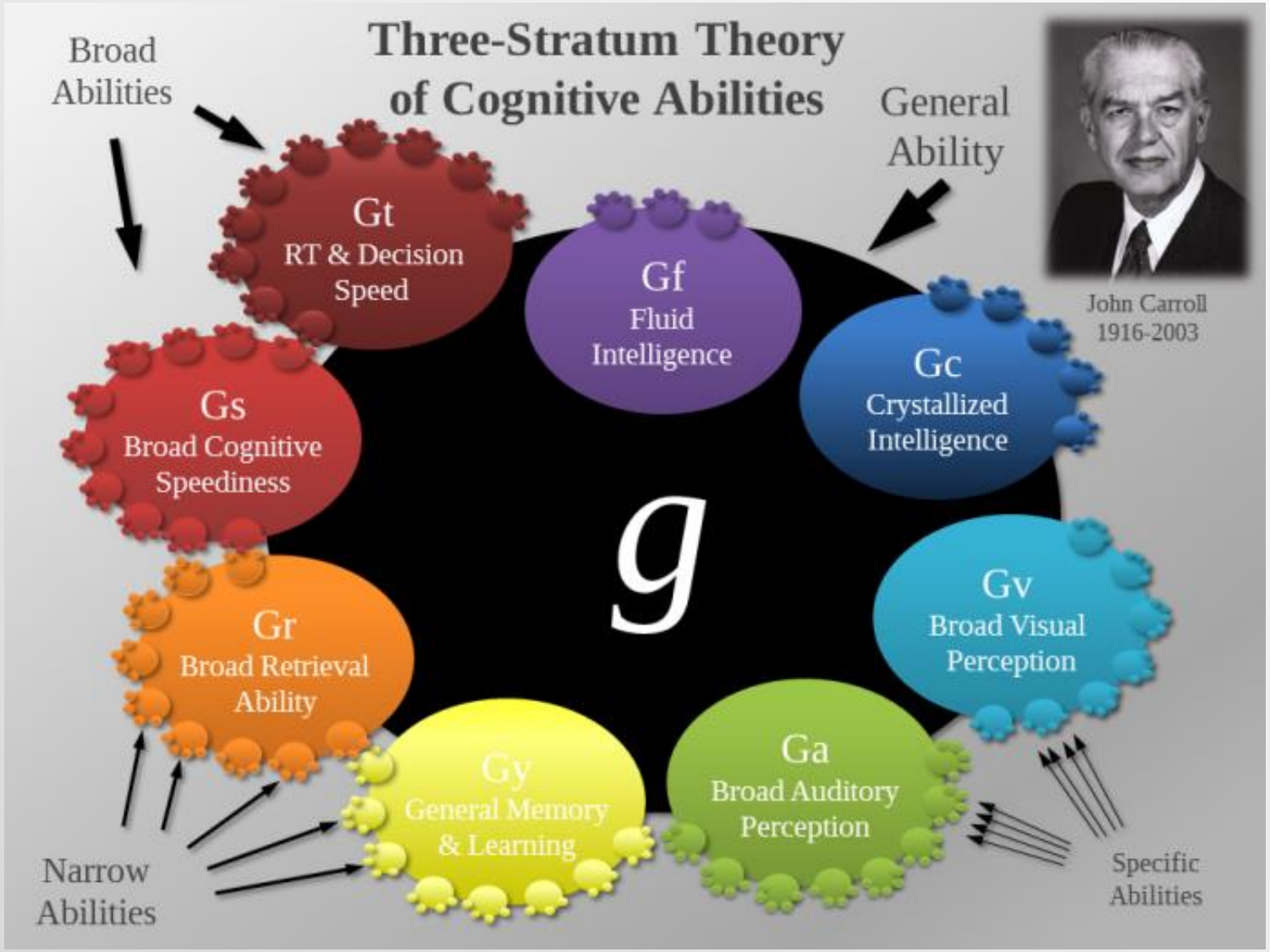
Măsurarea inteligenței & IQ

- 3 straturi



J. B. Carroll (1993), *Human cognitive abilities: A survey of factor-analytic studies*, Cambridge University Press, New York, NY, USA. [Grafic]

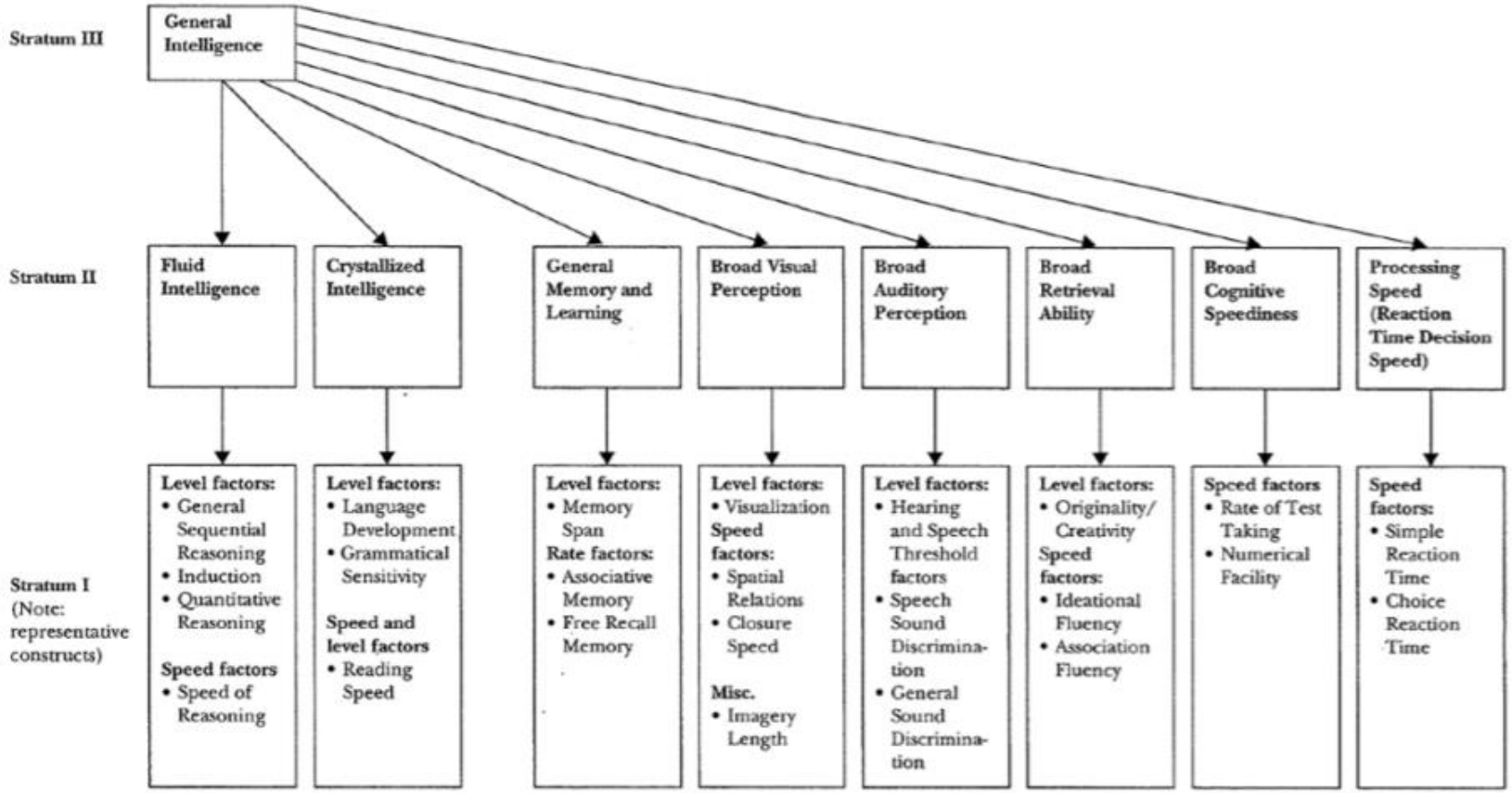
Inteligențe multiple



J. B. Carroll (1993), *Human cognitive abilities: A survey of factor-analytic studies*, Cambridge University Press, New York, NY, USA. [Grafic]

SYSTEMS LABORATORY

Carroll's Three-Stratum Theory of Cognitive Ability



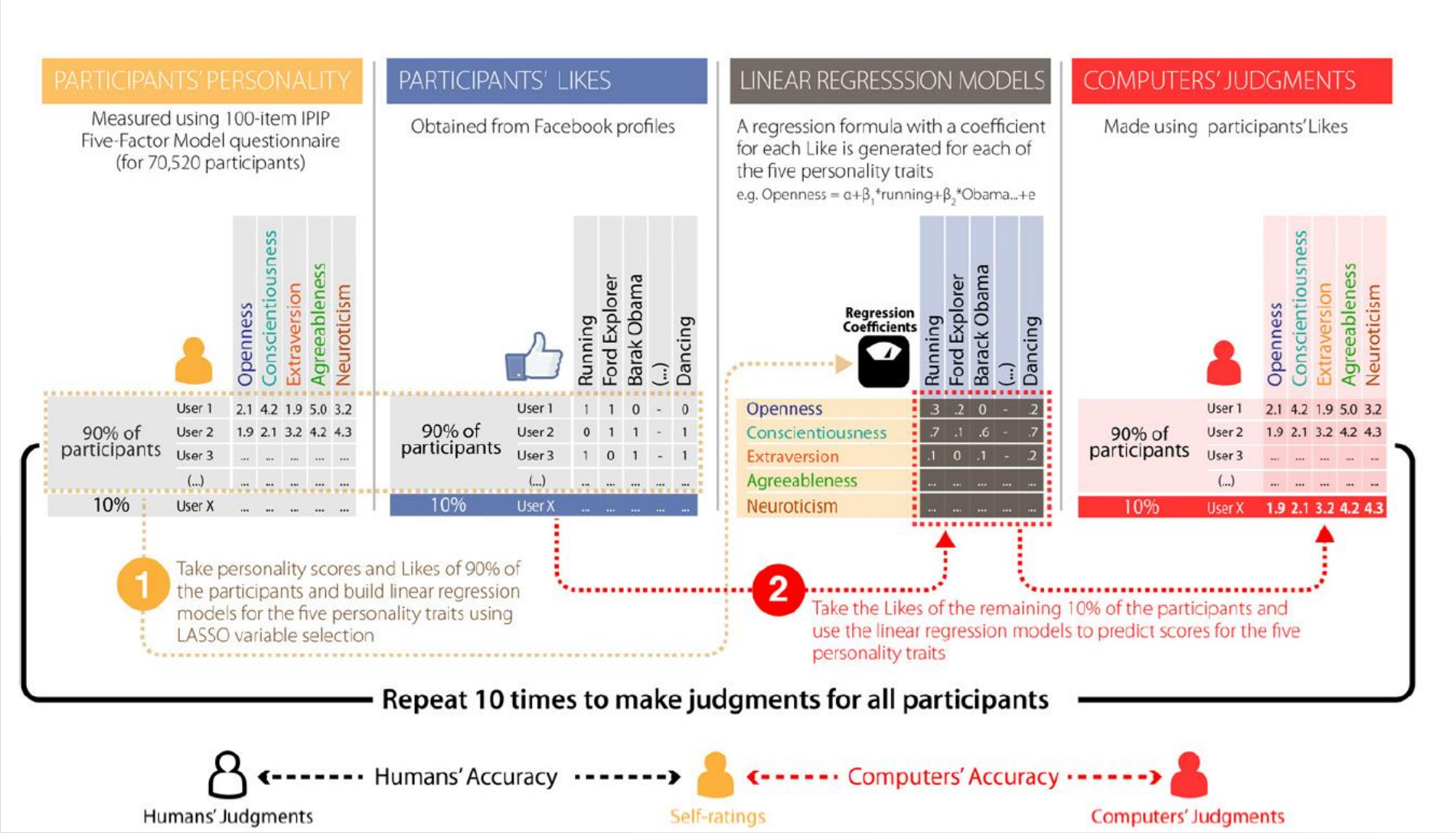
Adapted with permission from Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press. (p. 626).

6

PR network crunch

McDaniel, 2018

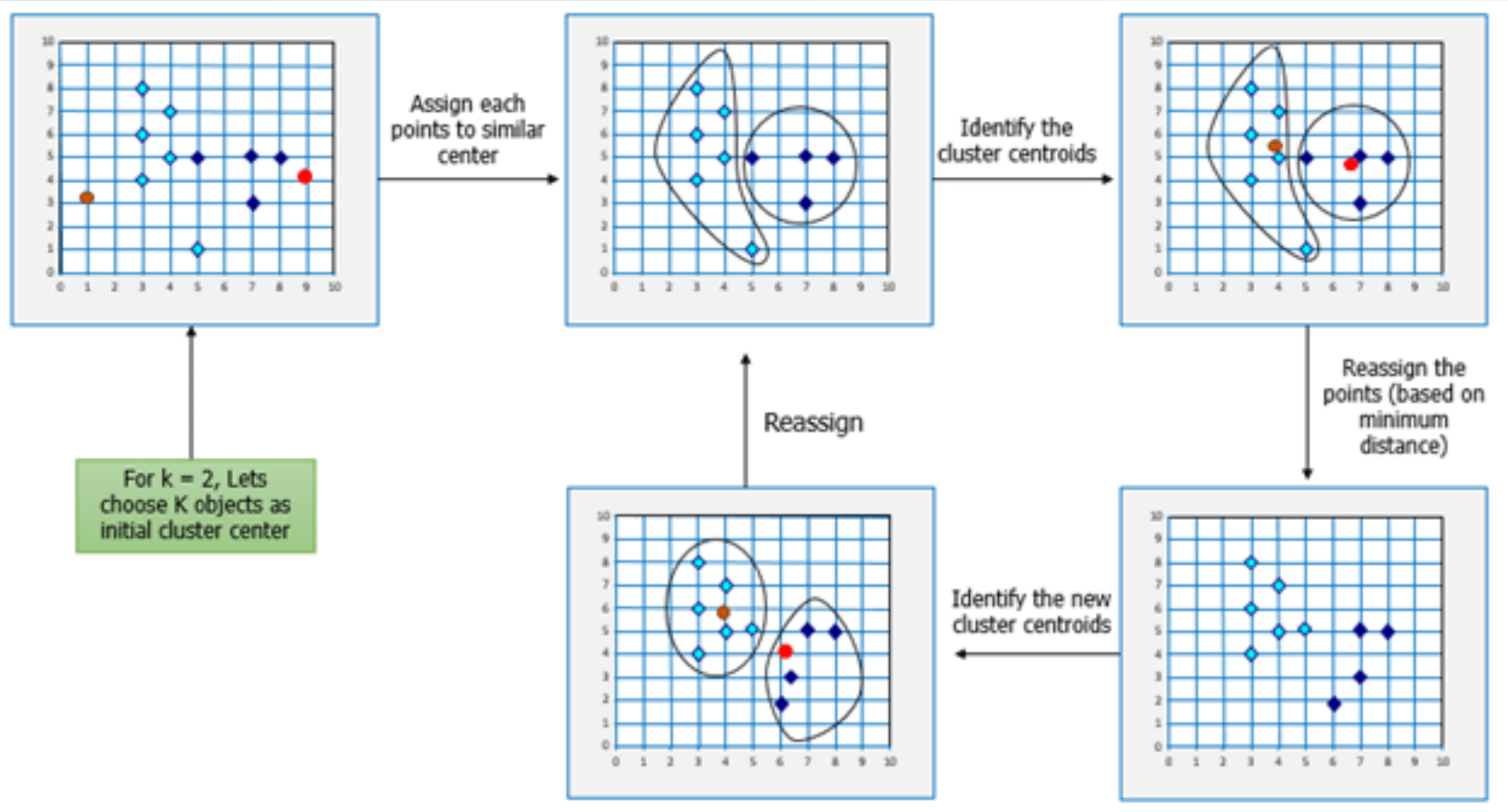
OCEAN și digital prediction



3.1.2 Analiza cluster

Tipologii

Analiza cluster K-means



The structure of videogame preference

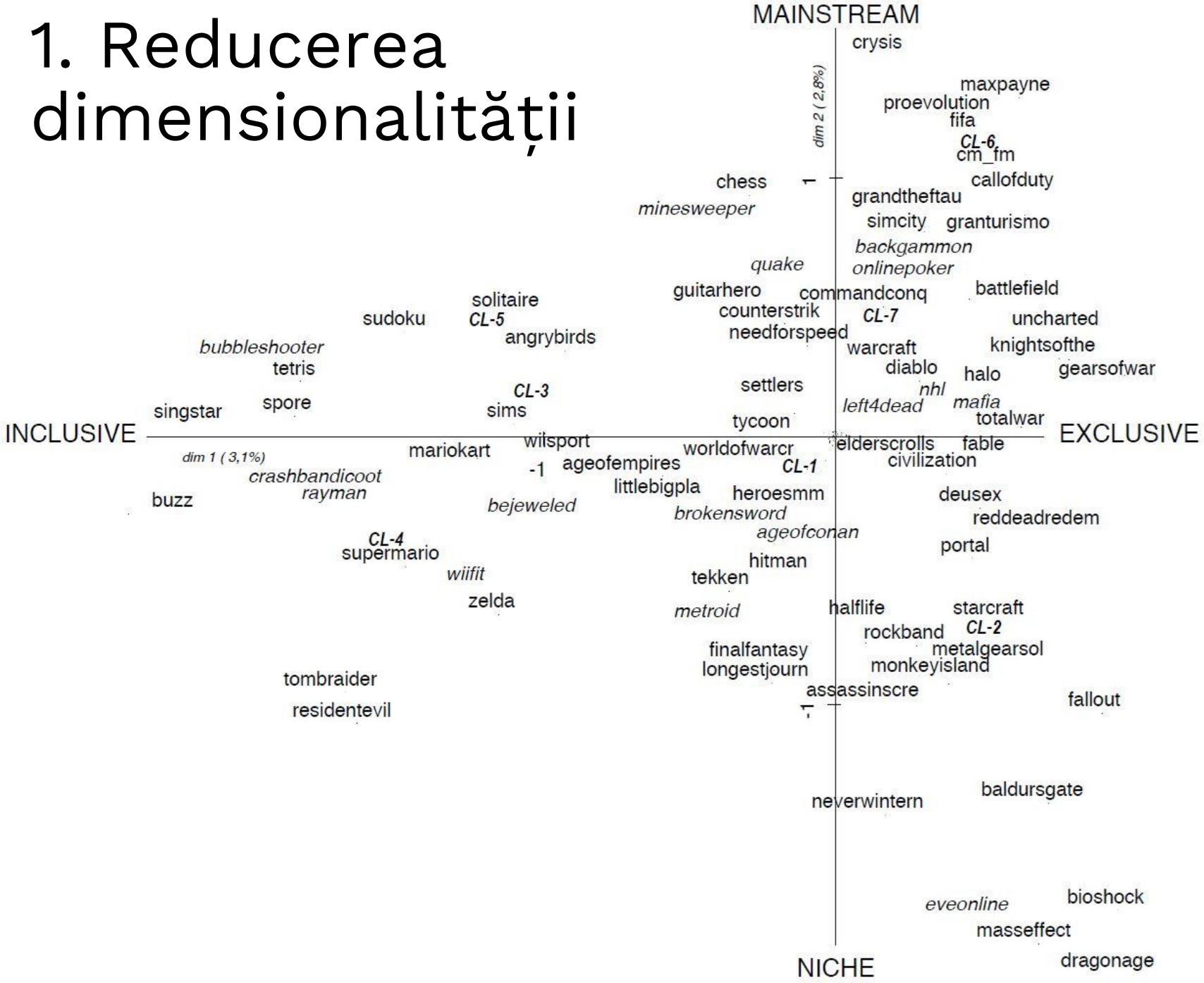
- Data: two broad surveys of Norwegian higher education students' cultural- media- and lifestyle preferences (2008, 2010)
- Of 2367 respondents, 758 students mentioned one or more favourite games, and 533 different titles were mentioned in total
  - Grouped in 328 game series, of which 64 chosen by 5+ respondents
  - Final dataset: a matrix consisting of 64 game series and 417 respondents
- Analyses:
  1. Dimensionality reduction: positioning games in a 2-dim. space
  2. Cluster analysis: identifying types of players

Klevjer & Hovden, 2017

Baza de date

	Joc 1	Joc 2	Joc 3	Joc 4	Joc 5	Joc 6	Joc 7	Joc 8	Joc 9
Resp 1	1		1	1					1
Resp 2	1		1	1					
Resp 3	1		1						
Resp 4		1			1	1	1		
Resp 5								1	1
Resp 6							1	1	1
Resp 7		1			1	1		1	
Resp 8	1		1		1				1
....									

1. Reducerea dimensionalității



Klevjer & Hovden, 2017

Reducerea dimensionalității

- Axele surprind **menționarea simultană** a jocurilor de respondenți
  - Jocurile menționate mai des simultan sunt considerate mai „apropiate”
- Două axe principale:
  - **Inclusiv / Exclusiv:**
    - „An opposition between an inclusive and family-friendly orientation at one end of the spectrum, and an exclusive, high-tech and dark/mature preference orientation at the other.”
  - **Mainstream / De nișă:**
    - „A niche-oriented or "geeky" taste at the lower end (although admittedly a large niche), versus a mainstream-oriented taste at the top”

Interpretarea axelor – ce etichete?

- Horizontal axis („inclusive / exclusive”):
  - On the **left** we find broadly popular classic series like Super Mario and The Sims, socially- and family-oriented series like Wii Sports, Singstar and Buzz, puzzle games like Tetris and Sudoku — and Angry Birds, which was at the peak of its popularity during the period of the surveys.
  - Around the **centre** and towards the **right** we find games that are generally more time-demanding.
- Vertical axis (mainstream / niche):
  - The **lower end** is dominated by fantasy and Science-Fiction of a particular flavour (e.g. Mass Effect, Dragon Age, Bioshock, Fallout, Zelda, Final Fantasy)
  - The **top end** is instead populated by distinctively mainstream and cross-media action fare: racing (Gran Turismo), shooting (Call of Duty, Crysis, Battlefield) and football (Championship Manager, Fifa, Pro Evolution Soccer)

2. Analiza cluster: gruparea respondenților

1: Strategists	2: Roleplayers	3: Partygamers	4: Nintendos	5: Casuals	6: Lads	7: Esporters
Sims 30%	Fallout 21%	Guitar Hero 43%	Zelda 53%	Angry Birds 53%	FIFA 45%	WoW 28%
Civilization 17%	Final Fantasy 16%	Singstar 25%	Super Mario 42%	Solitaire 24%	Call Of Duty 42%	Counterstrike 23%
Monkey Island 15%	Elder Scrolls 13%	Buzz 21%	Mario Kart 18%	Tetris 24%	GTA 34%	Starcraft 18%
Heroes M&M 12%	Mass Effect 13%	Fable 18%	Wii Sports 18%	Sims 20%	ChampM/FM 28%	ChampM/FM 14%
Age of Empires 11%	Assassins Creed 12%	FIFA 14%	Resident Evil 18%	Super Mario 9%	Halo 11%	Diablo 14%
ChampM/FM 10%	Red Dead Rev. 10%	Sims 11%	Sims 11%	Chess 9%	Gran Turismo 8%	Call Of Duty 10%

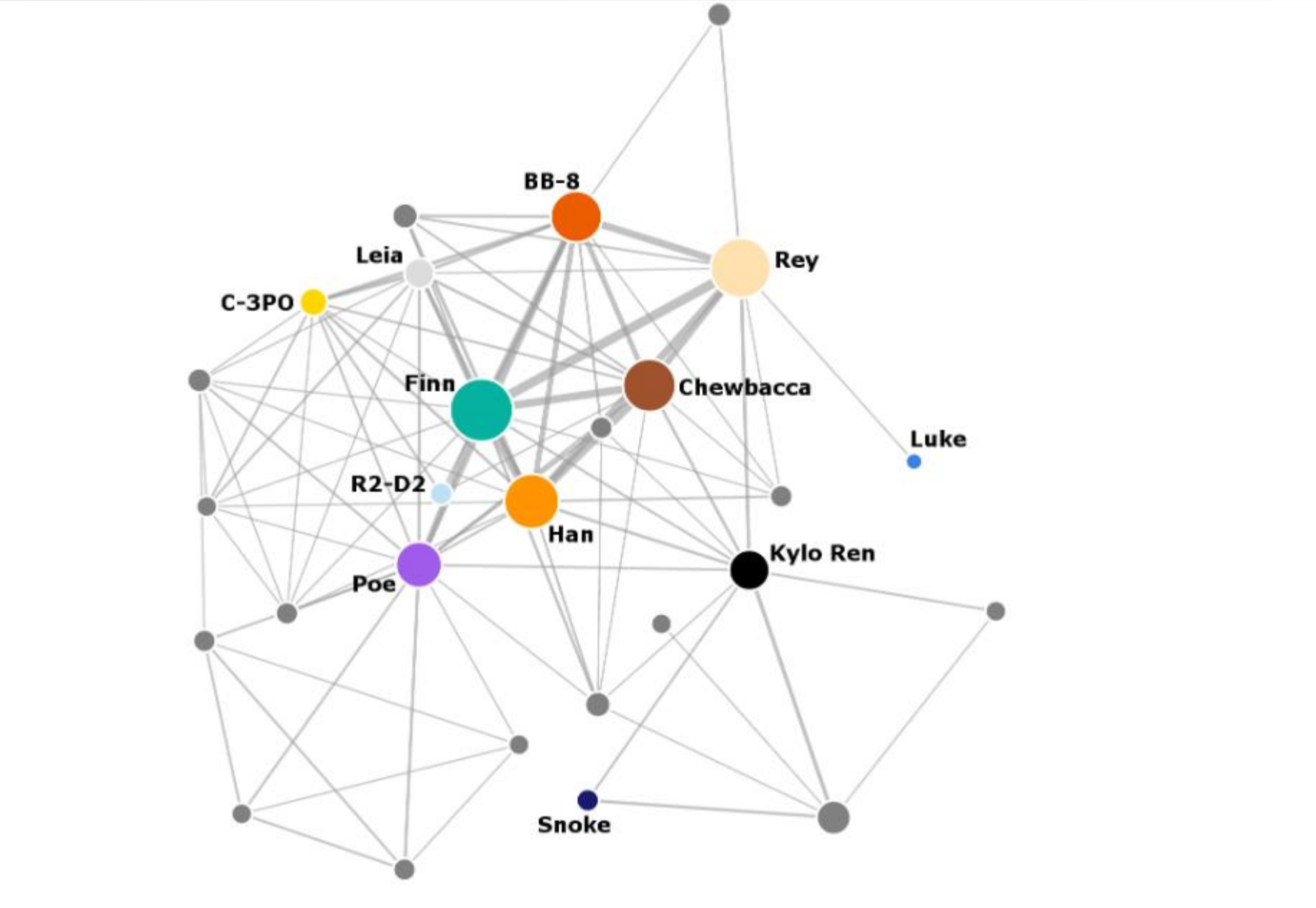
SYSTEMS LABORATORY

nch

	N=	mean age	% females	% play 5 d/ week+	% play weekly	% "very interested"
1: Strategists	110	24,4	46 %	18 %	47 %	18 %
2: Roleplayers	130	24,4	16 %	38 %	72 %	41 %
3: Partygamers	33	24,2	55 %	21 %	45 %	12 %
4: Nintendos	41	23,7	54 %	20 %	43 %	27 %
5: Casuals	49	26,1	69 %	17 %	32 %	10 %
6: Lads	75	24,9	4 %	47 %	73 %	40 %
7: Esporters	119	24,5	17 %	40 %	68 %	32 %
"Not interested"	1114	28,2	83 %	0 %	1 %	0 %
Total	**1671**	**27**	**66 %**	**11 %**	**20 %**	**10 %**

3.1.3 Analiza de rețea

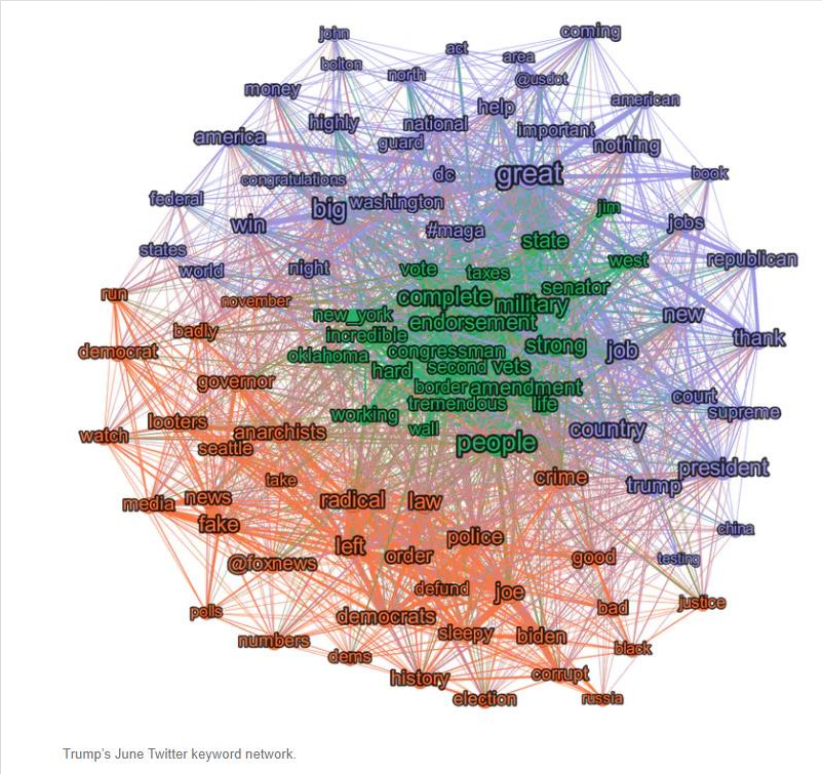
Centralitate, densitatea legăturilor


BB-8
Leia
Rey
C-3PO
Finn
Chewbacca
Luke
R2-D2
Han
Kylo Ren
Poe
Snoke

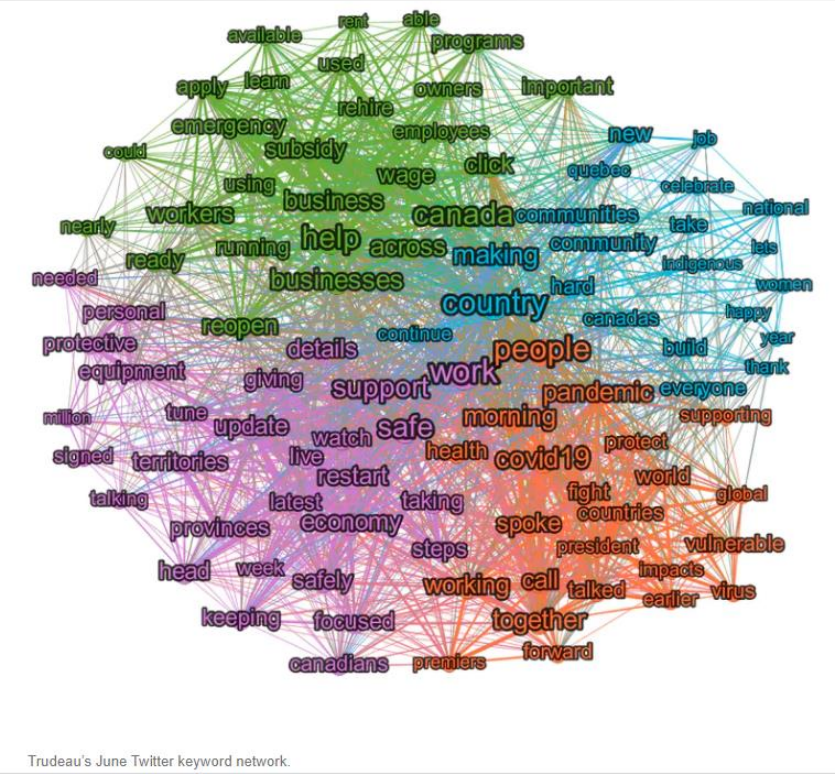
Interactiv

Rețele de co-ocurență pe Twitter

Trump, June 2020



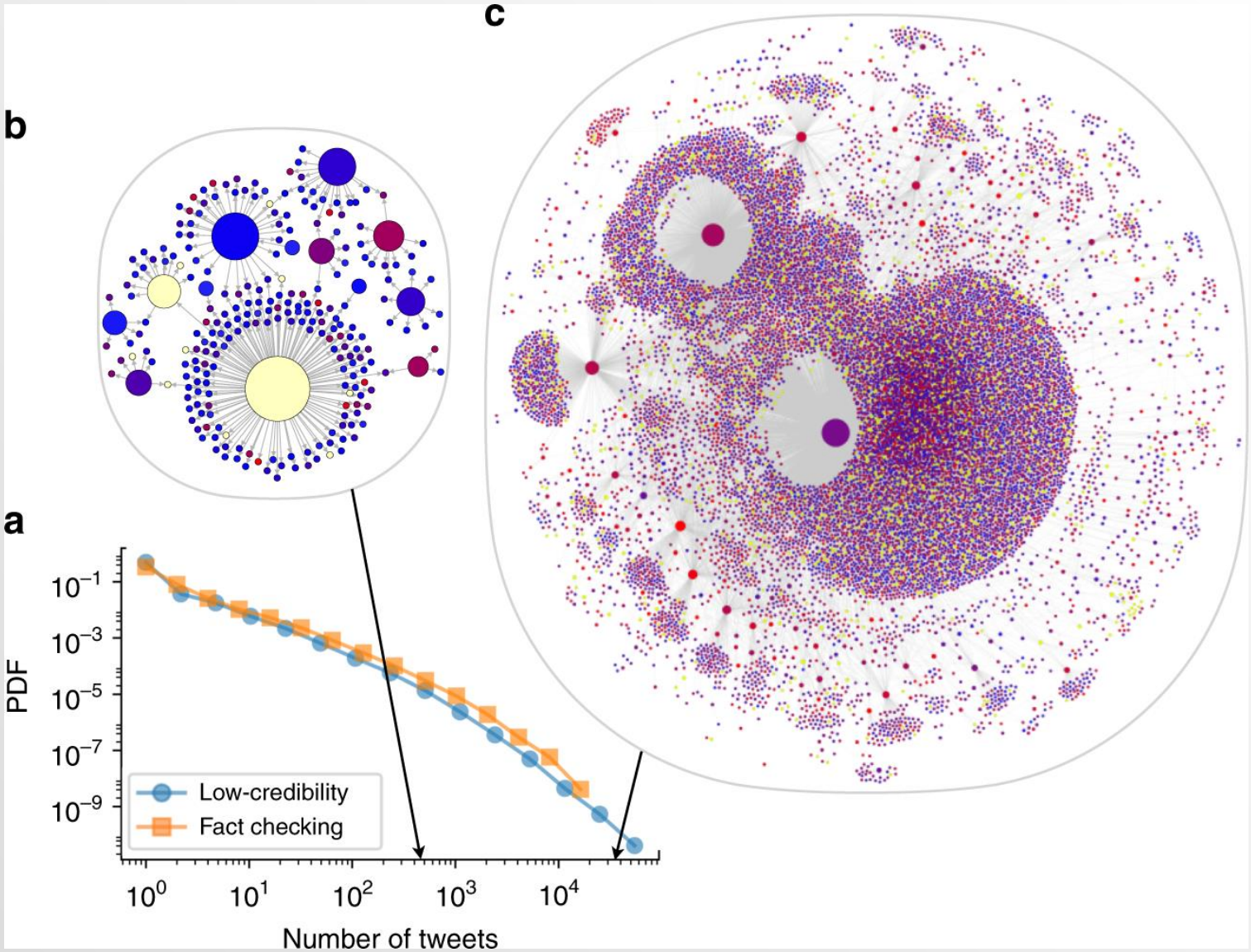
Trudeau, June 2020



Bonato & Nazareth, 2020

Bots in fake news retweet networks

- As illustrations, the diffusion networks of two stories are shown: **b** a medium-virality misleading article titled “FBI just released the Anthony Weiner warrant, and it proves they stole election”, published a month after the 2016 US election and **shared in over 400 tweets**; and **c** a highly viral fabricated news report titled “Spirit cooking”: Clinton campaign chairman practices bizarre occult ritual, published 4 days before the 2016 US election and **shared in over 30,000 tweets.**
- In both cases, only the largest connected component of the network is shown. Nodes and links represent Twitter accounts and retweets of the article, respectively. Node size indicates account influence, measured by the number of times an account was retweeted. **Node color represents bot score, from blue (likely human) to red (likely bot)**; yellow nodes cannot be evaluated because they have either been suspended or deleted all their tweets.



Chengcheng et al., in Nature, 2019

3.1.4 Analiza de regresie

Regresie și cauzalitate

Regresie

- Prezicem valori numerice continue (necunoscute) pe baza corelației cu alte variabile cunoscute
  - Cât costă un apartament cu 3 camere din centru (preț vs. zonă)?
  - Cu cât crește proporția de premii Nobel în funcție de consumul de ciocolată?

The NEW ENGLAND JOURNAL of MEDICINE

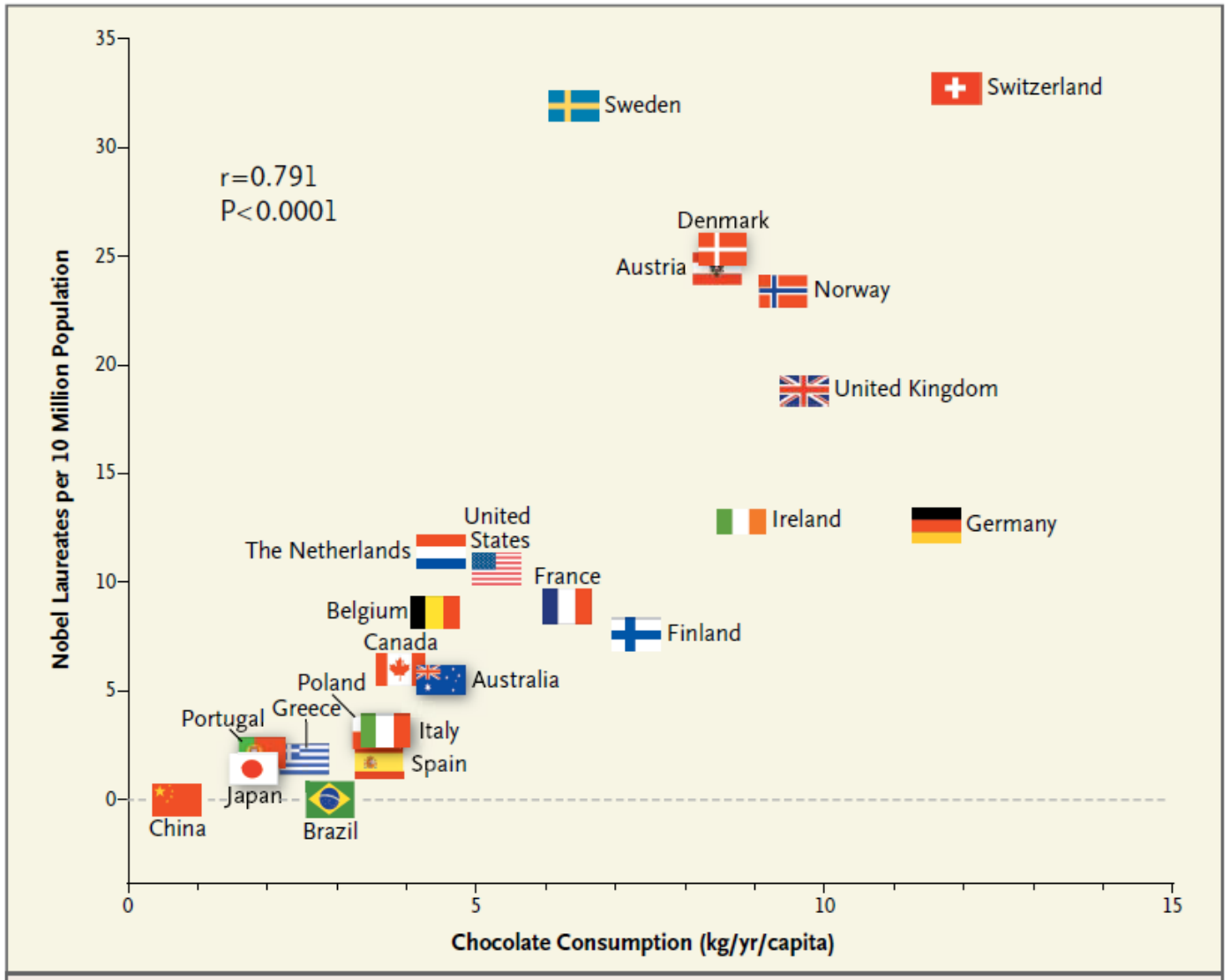
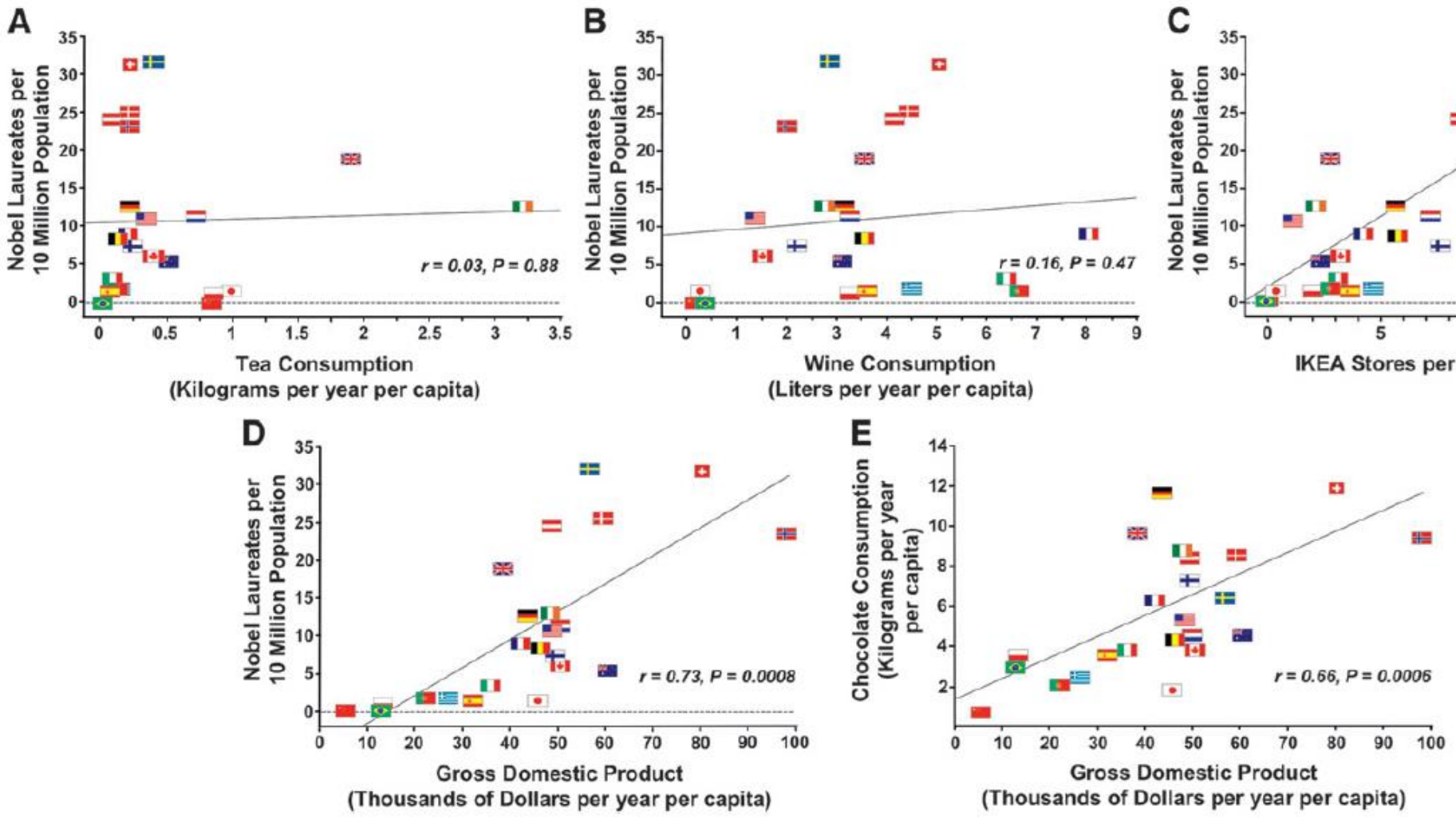
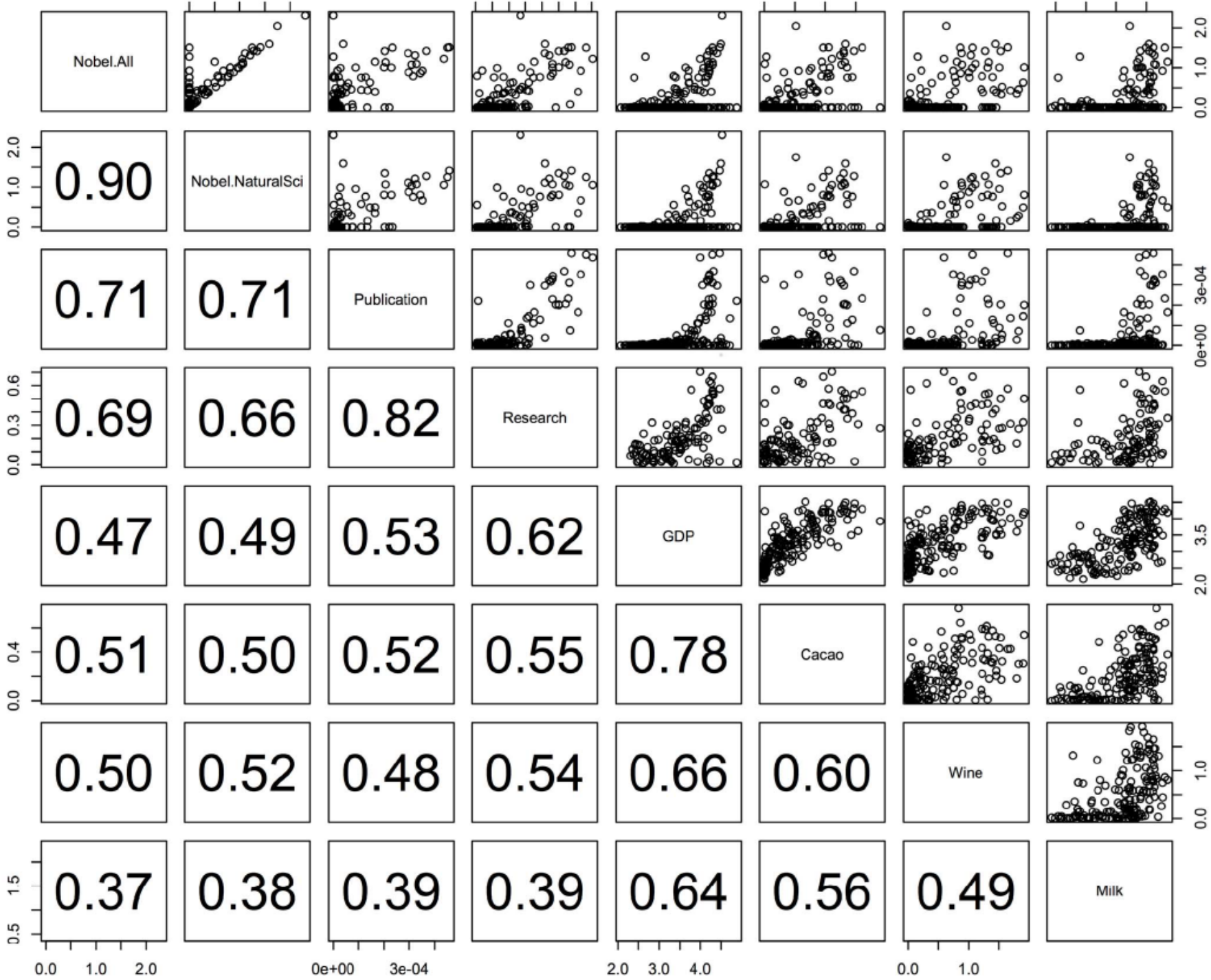


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.


A
Nobel Laureates per 10 Million Population
r = 0.03, P = 0.88
Tea Consumption (Kilograms per year per capita)
B
Nobel Laureates per 10 Million Population
r = 0.16, P = 0.47
Wine Consumption (Liters per year per capita)
C
Nobel Laureates per 10 Million Population
r = 0.82, P < 0.0001
IKEA Stores per 10 Million Population
D
Nobel Laureates per 10 Million Population
r = 0.73, P = 0.0008
Gross Domestic Product (Thousands of Dollars per year per capita)
E
Chocolate Consumption (Kilograms per year per capita)
r = 0.66, P = 0.0006
Gross Domestic Product (Thousands of Dollars per year per capita)


Nobel.All
Nobel.NaturalSci
Publication
Research
GDP
Cacao
Wine
Milk
0.90
0.71 0.71
0.69 0.66 0.82
0.47 0.49 0.53 0.62
0.51 0.50 0.52 0.55 0.78
0.50 0.52 0.48 0.54 0.66 0.60
0.37 0.38 0.39 0.39 0.64 0.56 0.49
0.0 1.0 2.0
0e+00 3e-04
2.0 3.0 4.0
0.0 1.0
0.0 1.0 2.0
0.0 0.3 0.6
2.0 3.5
0.0 0.4
0.0 1.0
0.5 1.5

3.1.5 Serii de timp

Evoluție și cauzalitate

Serii de timp

- Prezicem valori numerice continue (necunoscute) pe baza evoluției lor până acum
  - Câte grade vor fi mâine?
  - Care va fi prețul unui Bitcoin peste două zile?
- Combină
  - Proprietăți temporale interne: cicluri, tendințe
  - Modele explicative: cauze externe

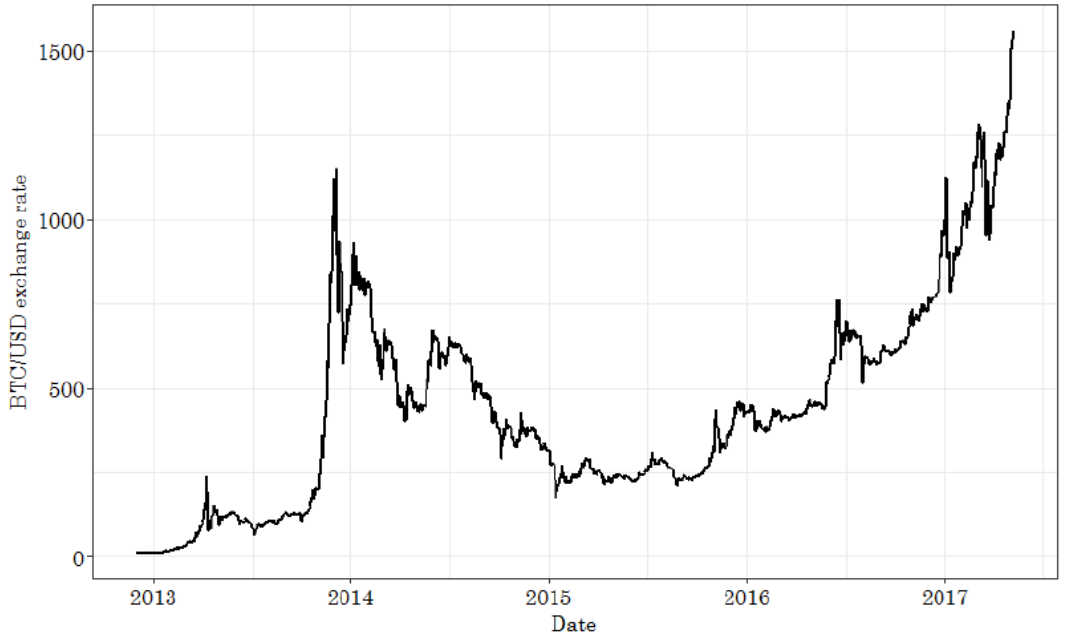


Figure 2: Bitcoin exchange rate with USD

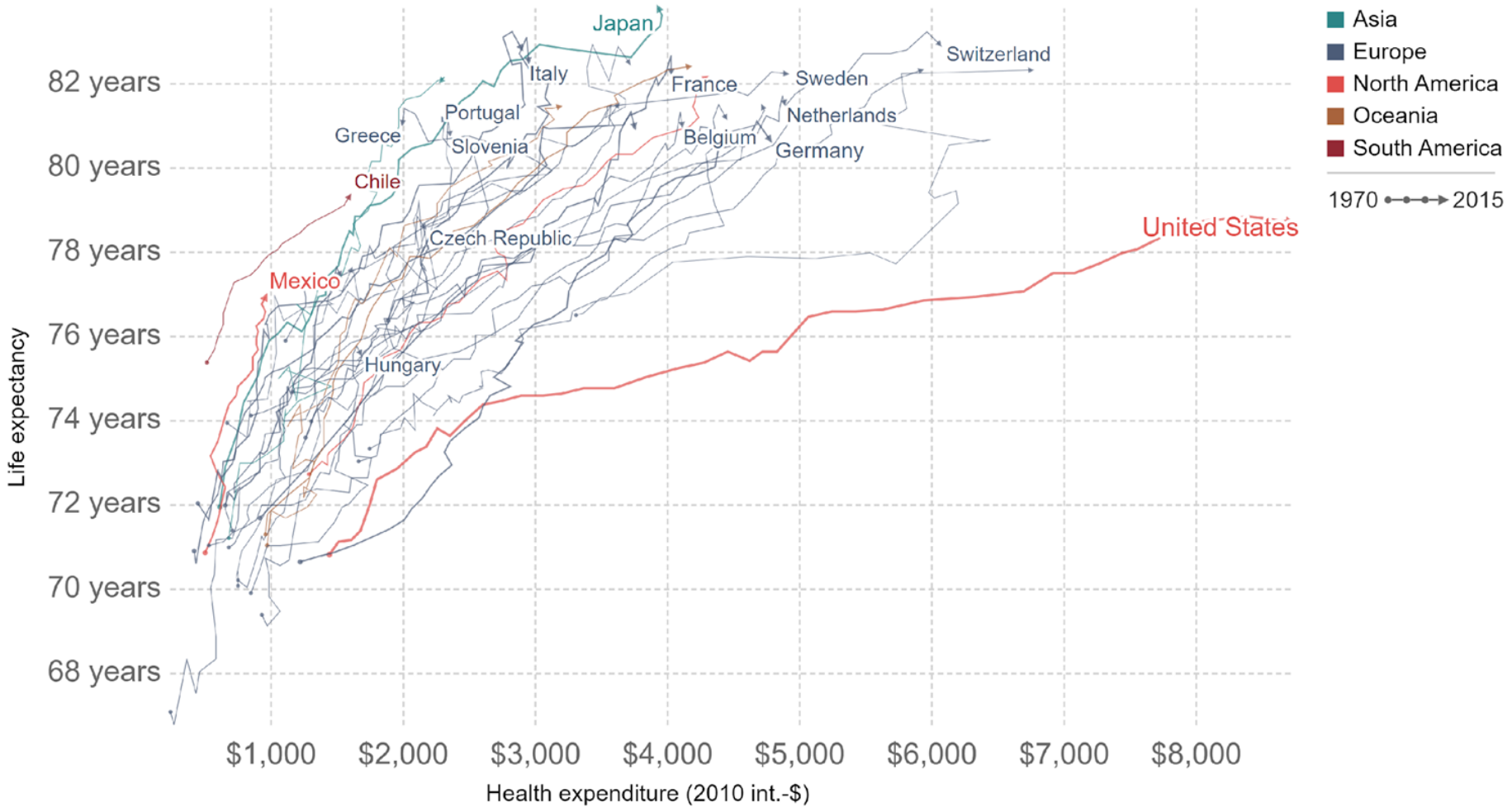
Poyser 2018

SYSTEMS LABORATORY

Life expectancy vs. health expenditure, 1970 to 2015

Our World in Data

Health financing is reported as the annual per capita health expenditure and is adjusted for inflation and price level differences between countries (measured in 2010 international dollars).



Source: World Bank, Health Expenditure and Financing - OECDstat (2017), Population (Gapminder, HYDE(2016) & UN (2019)), Our World In Data

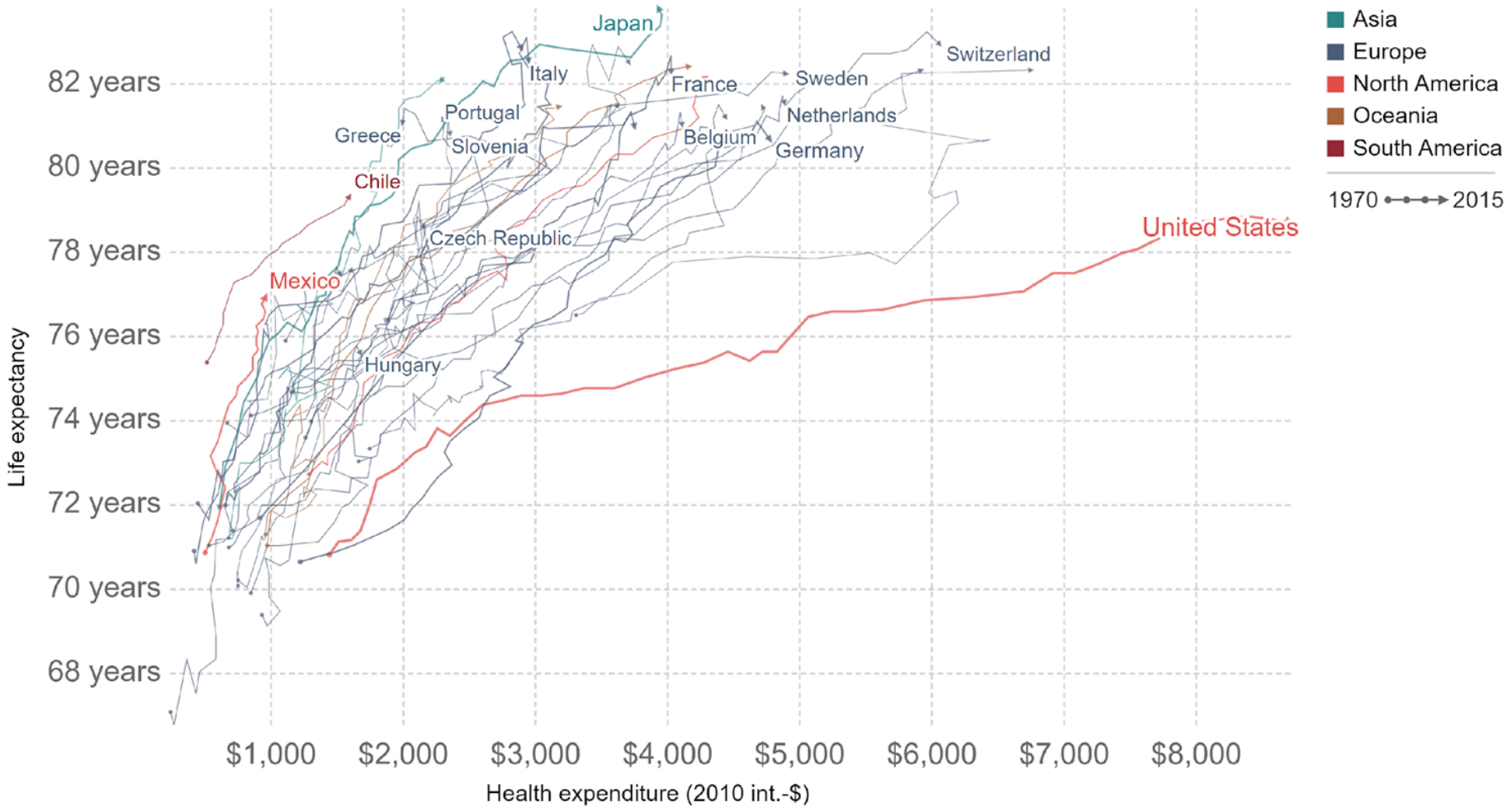
OurWorldInData.org/the-link-between-life-expectancy-and-health-spending-us-focus • CC BY

PR network crunch

Sursă

Life expectancy vs. health expenditure, 1970 to 2015

Health financing is reported as the annual per capita health expenditure and is adjusted for inflation and price level differences between countries (measured in 2010 international dollars).



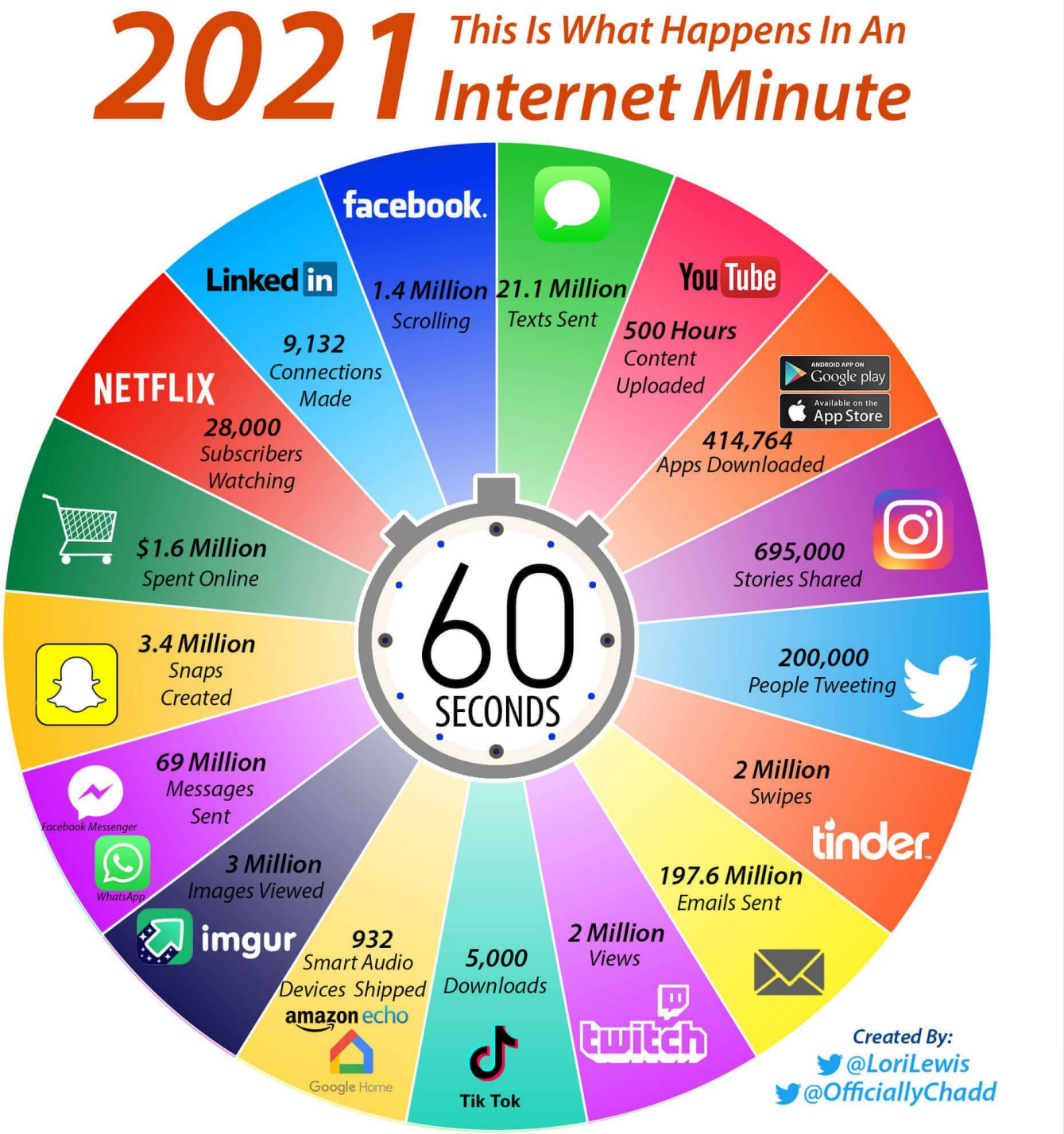
Source: World Bank, Health Expenditure and Financing - OECDstat (2017), Population (Gapminder, HYDE(2016) & UN (2019)), Our World In Data

OurWorldInData.org/the-link-between-life-expectancy-and-health-spending-us-focus • CC BY

#3.2 Capitalismul datelor și predicțiilor

Date + consolidare + Unique ID + Feedback loops

SYSTEMS LABORATORY

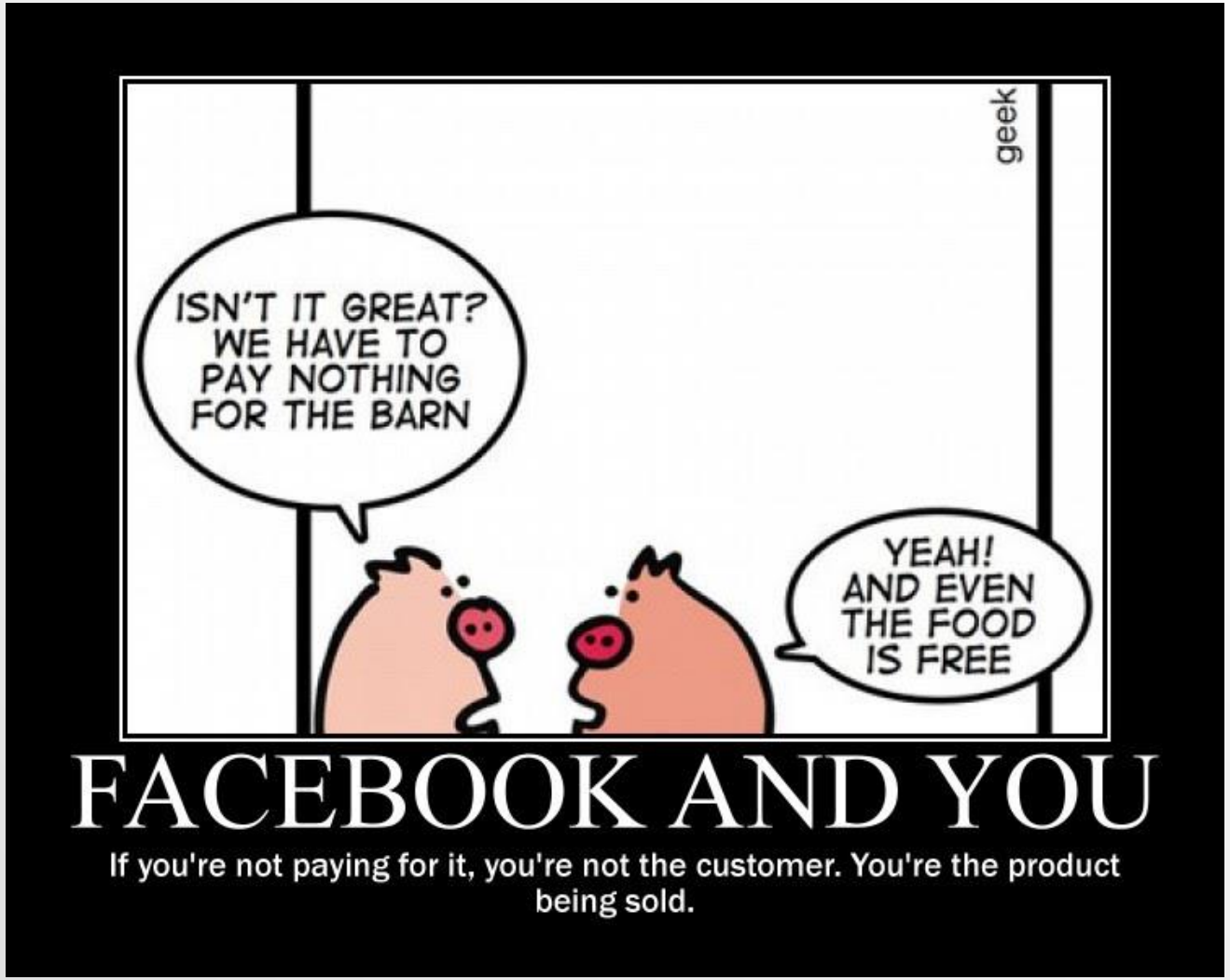


PR network crunch

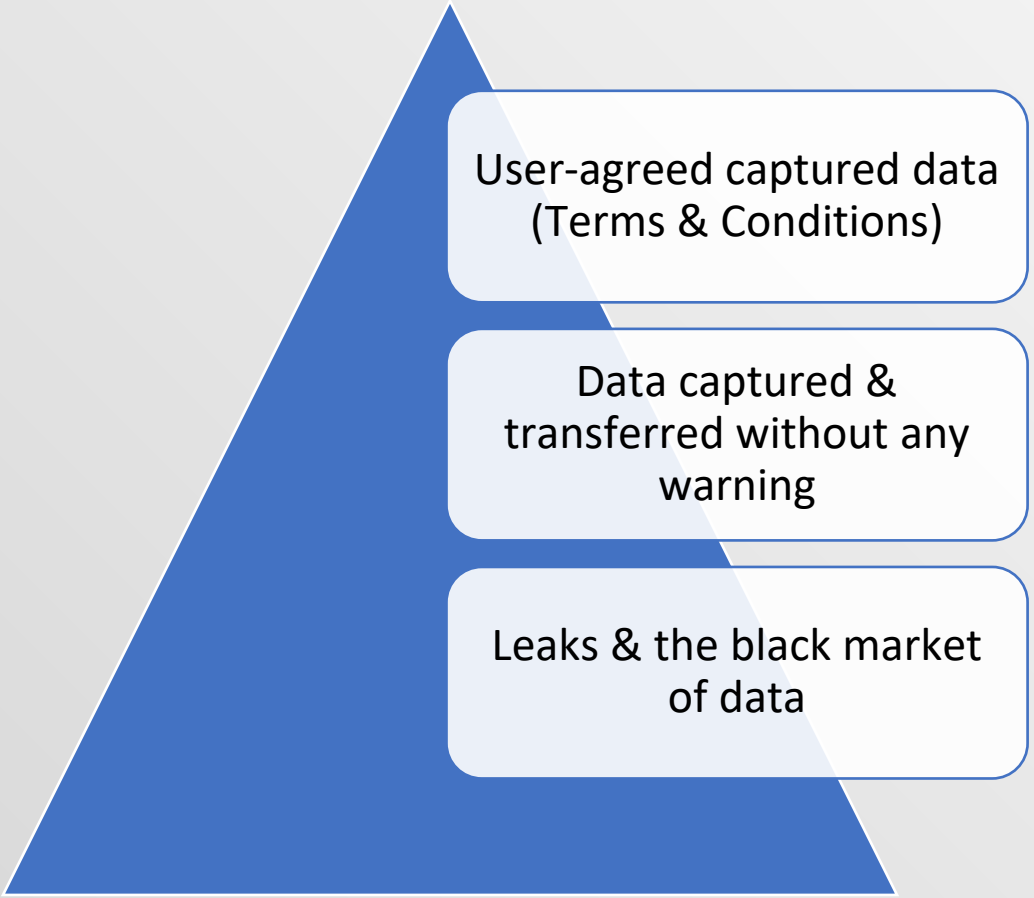
„Data is the new oil”

Capitalismul datelor

- Modelul de afaceri Google, Facebook etc. depinde de **marketingul algoritmic personalizat (AdTech)**
- **Piețele de date** sunt în expansiune

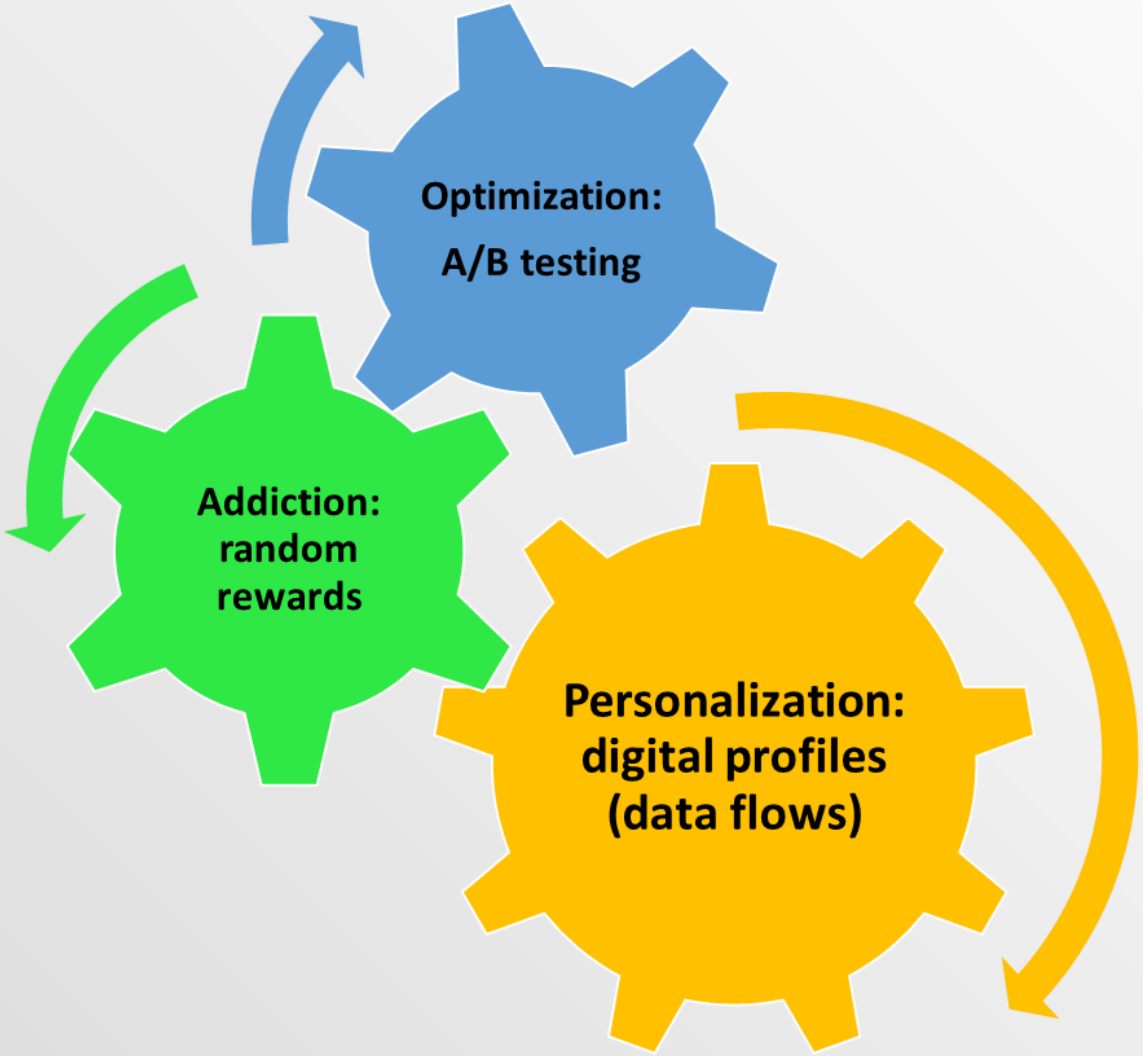


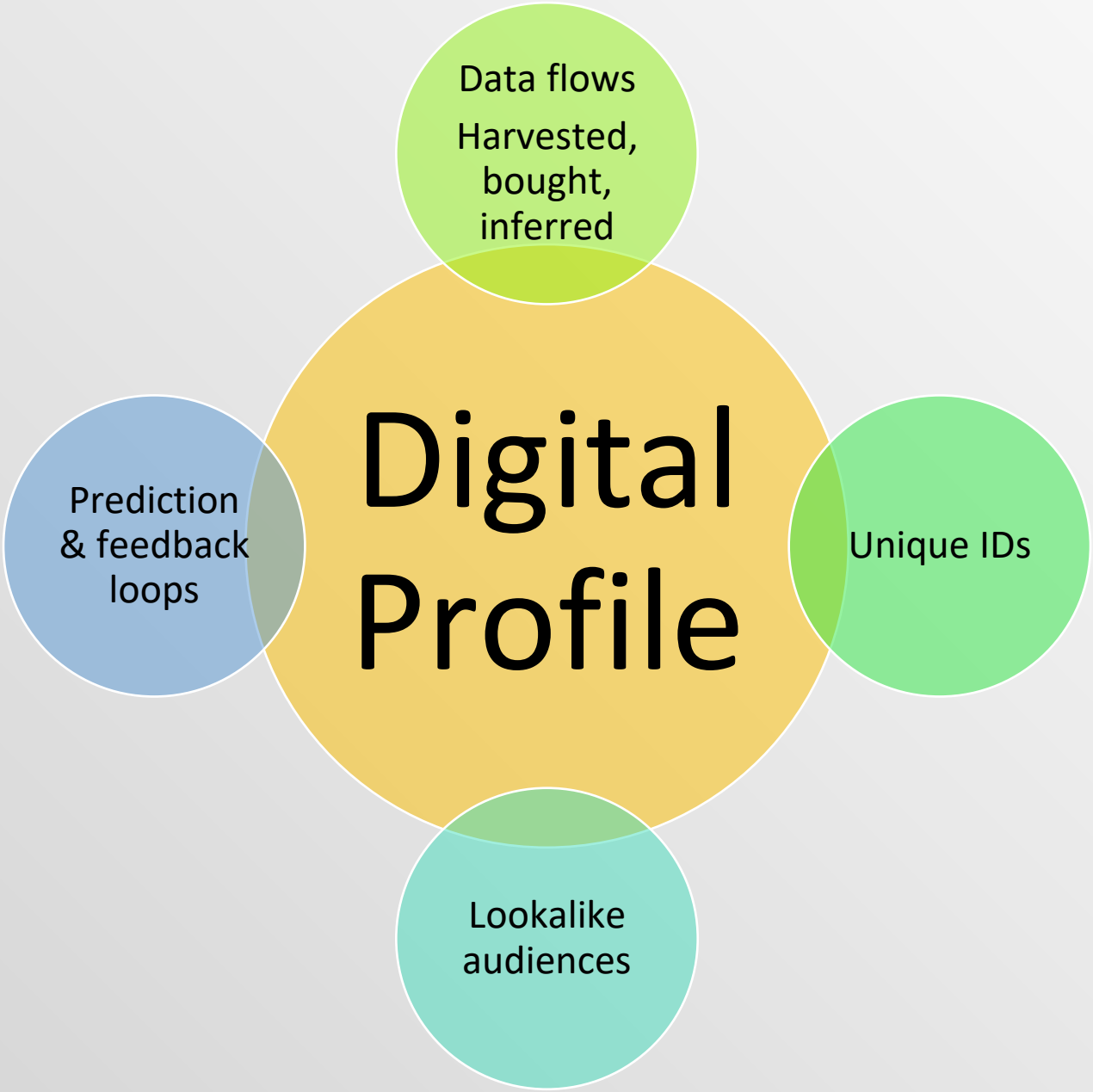
Controlul asupra datelor



Motoarele capitalismului datelor:

- Predicție prin A/B testing
- Predicție prin adicție
- Predicție prin **profiluri digitale & ML**




Data flows
Harvested, bought, inferred
Prediction & feedback loops
Digital Profile
Unique IDs
Lookalike audiences

Profiluri digitale

- Urme digitale procesate prin ML
  - Factuale și inferate
  - Cumpărate și consolidate
  - Cvasi-anonime: Unique ID
- Individualizate și colective
  - Ce facem noi
  - Ce fac cei din „audiențe similare” (lookalike audiences)

Apps

- Cele 10 aplicații studiate au transmis date spre 135 terțe părți implicate in AdTech
- Android Advertising ID a fost transferat spre 70 terțe părți
- Alte date transmise: adresa IP, locația GPS, gen, vârstă, diferite activități ale utilizatorilor inclusiv date foarte sensibile (din aplicații de dating sau sănătate)

OUT OF CONTROL

How consumers are exploited by the online advertising industry

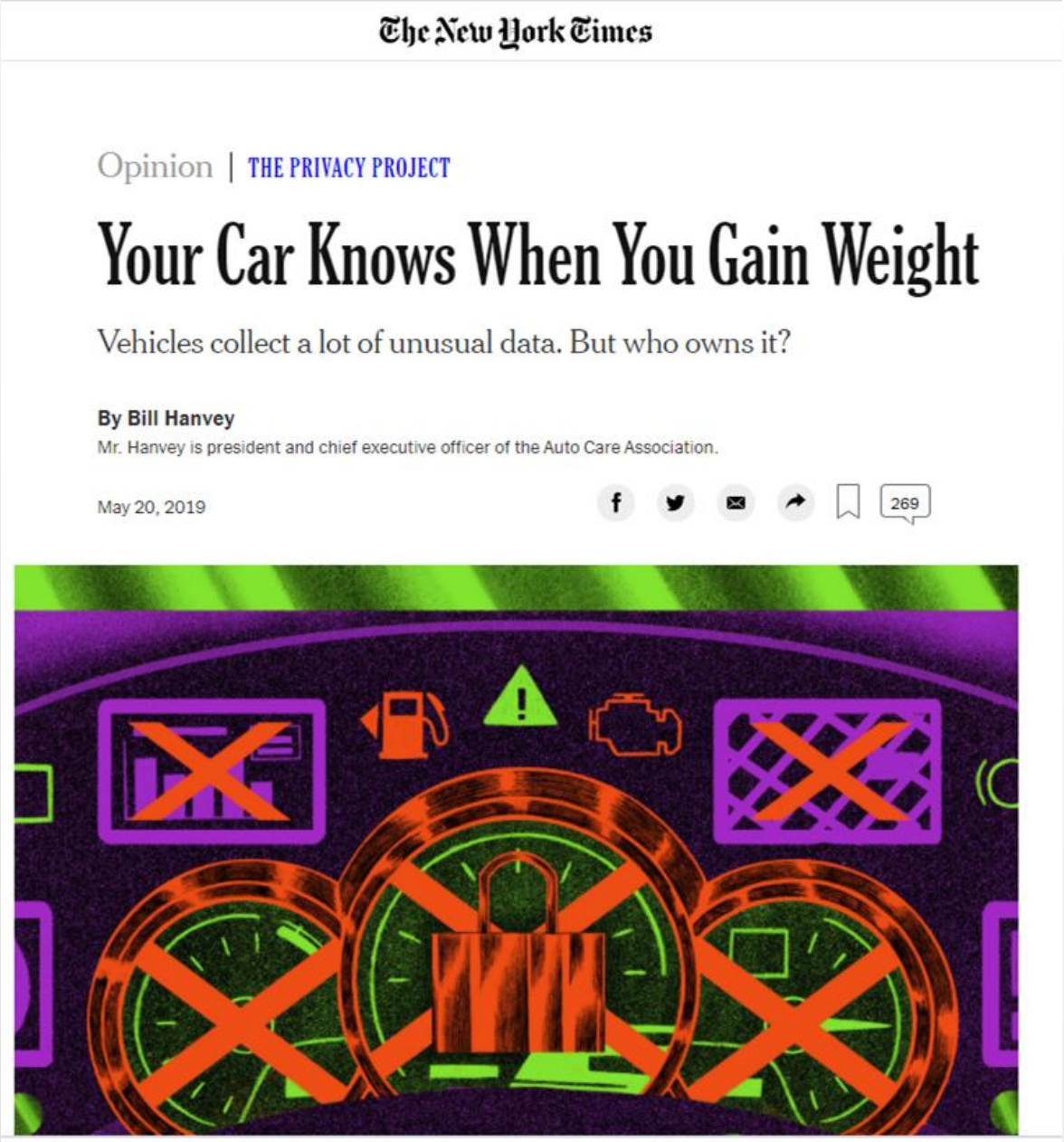
Norwegian Consumer Council 2020

App	Summary of findings
Clue	Sends birth year to **Amplitude**, **Apptimize**, and **Braze**. Sends Advertising ID to **Adjust**, **Amplitude**, and **Facebook**.
Grindr	Sends GPS coordinates to **AdColony**, **Braze**, **Bucksense**, **MoPub**, **OpenX**, **Smaato**, **PubNative**, **Vungle**, and others. Sends the IP address to **AppNexus** and **Bucksense**, and information about "relationship type" to **Braze**. Sends Advertising ID to all of these third parties and others, except **Braze.**
Happn	Sends country, gender and age segment of the user to **Google**. Sends Advertising ID to **Adjust** and **Facebook**.
Muslim: Qibla Finder	Sends IP address to **Appodeal**. Sends Advertising ID to **AppLovin**, **Appodeal**, **Facebook**, and **Liftoff**.
My days	Sends GPS coordinates and Wi-Fi access point information to **Neura**, **Placed**, and **Placer**. Sends IP address and a list of installed apps on the phone to **Placed**. Sends Advertising ID to **AppLovin**, **Liftoff**, **Google**, **Ogury Presage**, and **Placed**.
My Talking Tom 2	Sends IP address to **Mobfox**, **PubNative**, and **Rubicon Project**. Sends Advertising ID to **AppsFlyer**, **AppLovin**, **Facebook**, **IQzone**, **ironSource**, **Mobfox**, **Outfit7**, and **Rubicon Project**.
OkCupid	Sends GPS coordinates and answers to personal questions to **Braze**. Sends detailed device information to **AppsFlyer**. Sends Advertising ID to **AppsFlyer**, **Facebook** and **Kochava**.
Perfect365	Sends various location data such as GPS coordinates and Wi-Fi access point information to **Fysical**, **Safegraph**, and **Vungle**. Sends GPS coordinates unencrypted to **Receptiv**. Sends Advertising ID to **Amazon**, **Chocolate**, **Facebook**, **Fluxloop**, **Fyber**, **Fysical**, **InMobi**, **Inner-Active**, **Ogury Presage**, **Safegraph**, **Receptiv**, **Unacast**, **Unity3d**, and **Vungle**.
Tinder	Sends GPS position and "target gender" to **AppsFlyer** and **LeanPlum**. Sends Advertising ID to **AppsFlyer**, **Branch**, **Facebook**, and **Salesforce** (**Krux**).
Wave Keyboard	Sends Advertising ID to **Crashlytics**, **Facebook**, **Flurry**, **OneSignal**.

Norwegian Consumer Council 2020

Cars

- „Cars not only know how much we weigh but also track how much weight we gain. They know how fast we drive, where we live, how many children we have — even financial information. Connect a phone to a car, and it knows who we call and who we text.”
- „But who owns and, ultimately, controls that data? And what are carmakers doing with it?”



The New York Times

Opinion | THE PRIVACY PROJECT

Your Car Knows When You Gain Weight

Vehicles collect a lot of unusual data. But who owns it?

By Bill Hanvey

Mr. Hanvey is president and chief executive officer of the Auto Care Association.

May 20, 2019

269

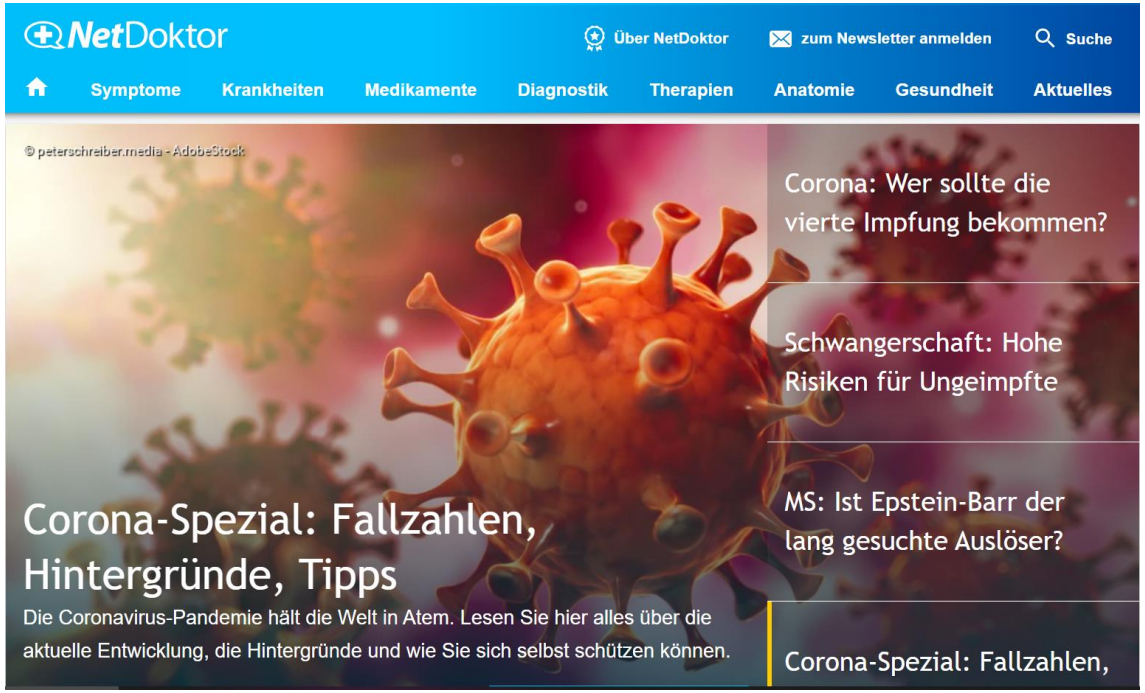
Hanvey 2019, NY Times The Privacy Project

1. The GDPR

General Data Protection Regulation

GDPR vs Google Analytics

- Dec. 2021: The use of Google Analytics on NetDoktor violates GDPR (Austrian Data Regulator) (Wired)
- Dec. 2021: European Parliament's Covid-19 testing website breached GDPR by using cookies from Google Analytics and Stripe (European Data Protection Regulator) (Wired)
- July 2020: Privacy Shield, use to transfer data from EU to US, is invalid (European Court of Justice) (Wired)
- Data are not protected from access by US intelligence agencies



EU-US Data Transfers

Since 2013 this case on US mass surveillance and EU companies feeding into it, is pending. ...

Read more

Forced Accounts

Many companies take advantage of the digitalisation of their services and collect far more data about their customers than necessary. ...

Read more

Forced Consent (DPAs in Austria, Belgium, France, Germany and Ireland)

When relying on consent as a legal basis to process personal data, companies need to comply with the stringent requirements contained in the GDPR. ...

Read more

Right to rectification (Art 16 GDPR)

There is a basic right for every user to request their data to be corrected in case of inaccuracies: quickly, easily, and free of charge. ...

Read more

Streaming Services

In 2018, data subjects represented by *noyb*, submitted requests to access information about their data with eight streaming services. ...

Read more

noyb

News Projects Resources

Current Enforcement Projects

Advertising IDs

We all use smart-phones. With them we surf the Internet, do research and download and use apps. ...

Read more

Cookie Banners

Cookie banners are regulated under the ePrivacy directive. ...

Read more

Credit Scoring

Credit scoring is the practice of giving individuals a credit worthiness score in order to determine whether to lend them money or extend a post-paid service ...

Read more

Data Portability (Art 20 GDPR)

Data portability empowers individuals to receive personal data they provided to the controller and, where technically feasible, to request this p ...

Read more

Dating Apps

Dating apps are gaining popularity, especially after Covid-19. ...

Read more

Encryption - Privacy by Design

Our emails always contain personal data. ...

Read more

Max Schrems

Max Schrems in 2016

Born	Maximillian Schrems October 1987 (age 34) Salzburg, Austria
Education	Law, University of Vienna
Occupation	Lawyer, author, privacy activist
Organization	NOYB – European Center for Digital Rights
Known for	Privacy activism

Privacy activism: noyb.eu
My privacy is none of your business

Buy a phone, get a tracker: unauthorized tracking code illegally installed on Android phones

As reported by the Financial Times, *noyb* launched further action against Google's AAID (Android Advertising Identifier), following similar complaints against Apple's IDFA. The somewhat hidden ID allows Google and all apps on the phone to track a user and combine information about online and mobile behaviour. While these trackers clearly require the users' consent (as known from "cookie banners"), Google neglects this legal requirement. *noyb* therefore filed a complaint against Google's tracking code AAID.

- Complaint filed to the Data Protection Authority of France (PDF)

Source

1. Amazon GDPR fine – €746 million

On July 16, 2021, the **Luxembourg National Commission for Data Protection** (**CNDP**) issued the **biggest fine ever for the violation of the GDPR** in the amount of **€746 million ($888 million)** to **Amazon.com** Inc.

The fine was issued as a result of a complaint filed by 10,000 people against Amazon in May 2018, through a French privacy rights group that **promotes and defends fundamental freedoms in the digital world- La Quadrature du Net.**

The CNPD opened an investigation into how Amazon processes personal data of its customers and found infringements regarding Amazons' advertising targeting system that was carried out without proper consent.

Source

La Quadrature du Net

ABOUT US

La Quadrature du Net (LQDN) promotes and defends fundamental freedoms in the digital world. We fight against censorship and surveillance, both from States or private companies. We questions how the digital world and society influence each other. We work for a free, decentralised and empowering Internet.

Contact **Follow LQDN**

Source

TOP 5 BIGGEST GDPR FINES

1	Amazon Europe	€746.000.000
2	WhatsApp	€225 000 000
3	Google Inc.	€50 000 000
4	H&M Hennes & Mauritz	€35 258 708
5	TIM - Telecom Provider	€27 800 000

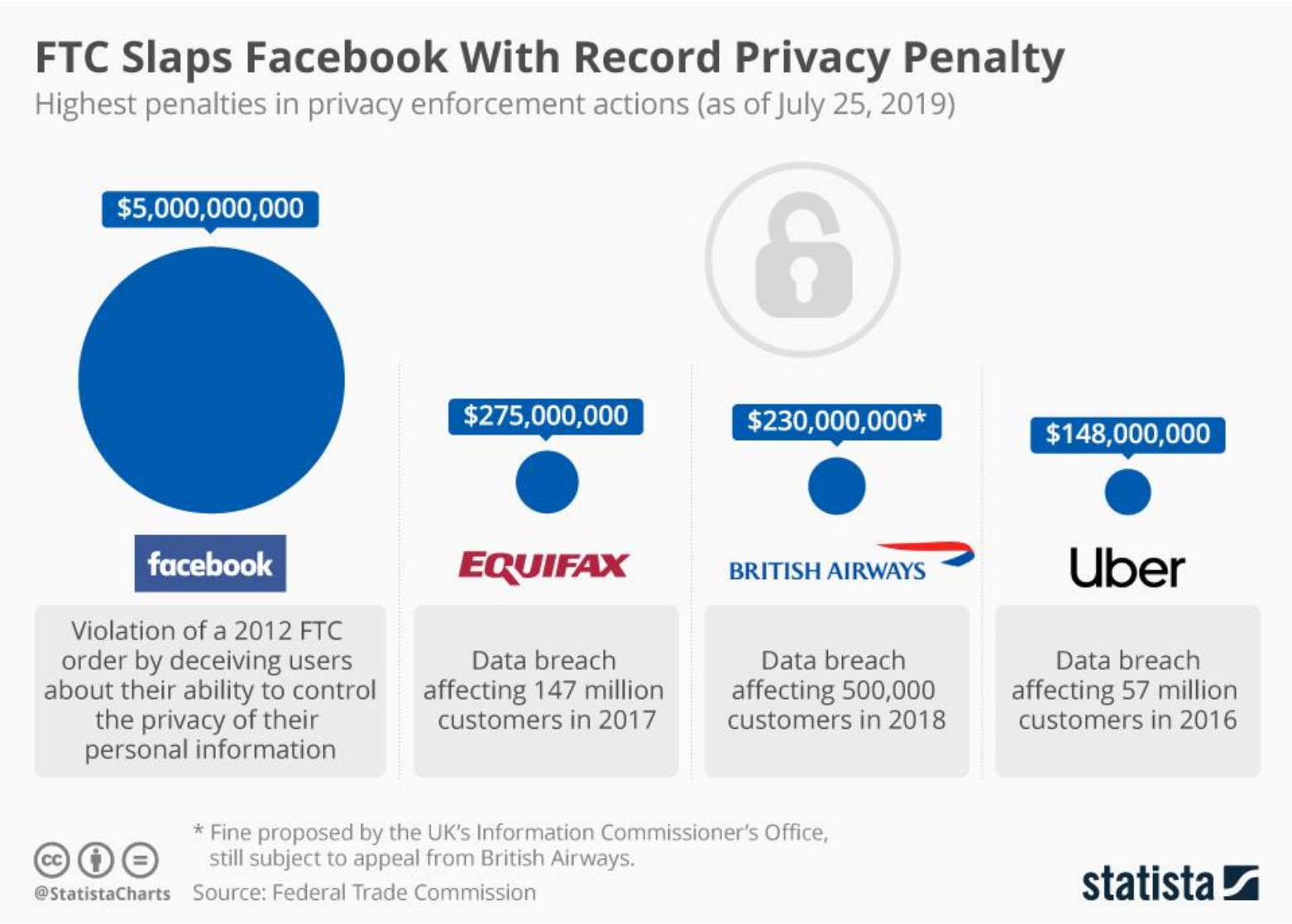
Source

Source

Federal Trade Commission privacy fines

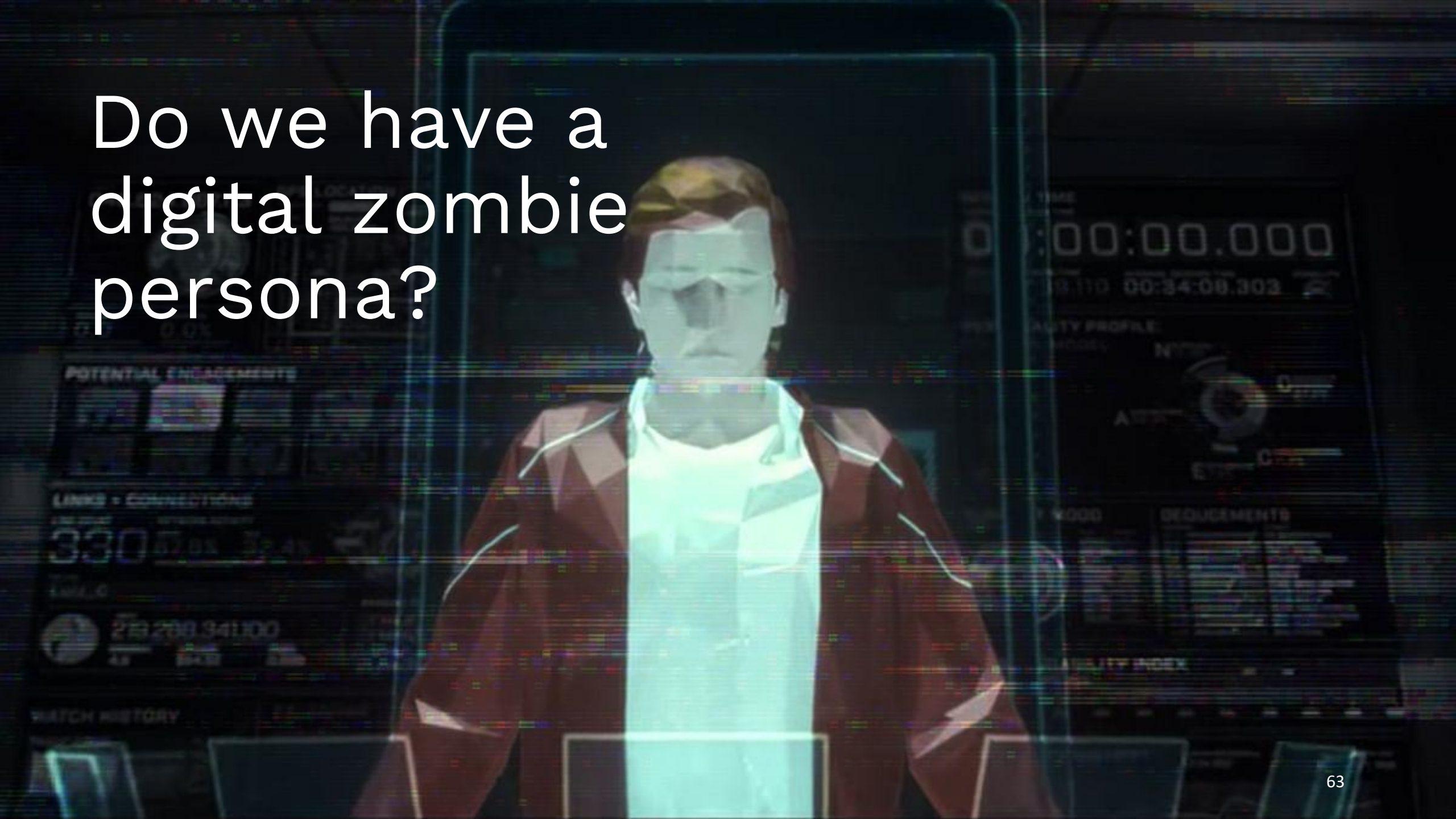
- Human rights (EU) vs. consumer choice (US)
- “Despite repeated promises to its billions of users worldwide that they could control how their personal information is shared, Facebook undermined **consumers’ choices**”

Source



Source

Do we have a digital zombie persona?



Frederike Kaltheuner
Privacy International report

- “More than 600 apps had access to my iPhone data” (BBC interview)
- “Over the course of a single week, Quantcast has amassed over 5300 rows and more than 46 columns worth of data including URLs, time stamps, IP addresses, cookies IDs, browser information and much more” (PI investigation)

64

Invasion story

- Datele sunt “bizar de specifice”
- Prezise/inferate: gender, age, number of children and their ages, education level, and gross yearly household income
- Travel and leisure to Canada
- Frequent transactions in Bagel Restaurants
- City Prosperity:World-Class Wealth
- Alcohol at Home Heavy Spenders
- Baby Nappies & Wipes
  - “wrong, very wrong”

(PI investigation)

SYSTEMS LABORATORY

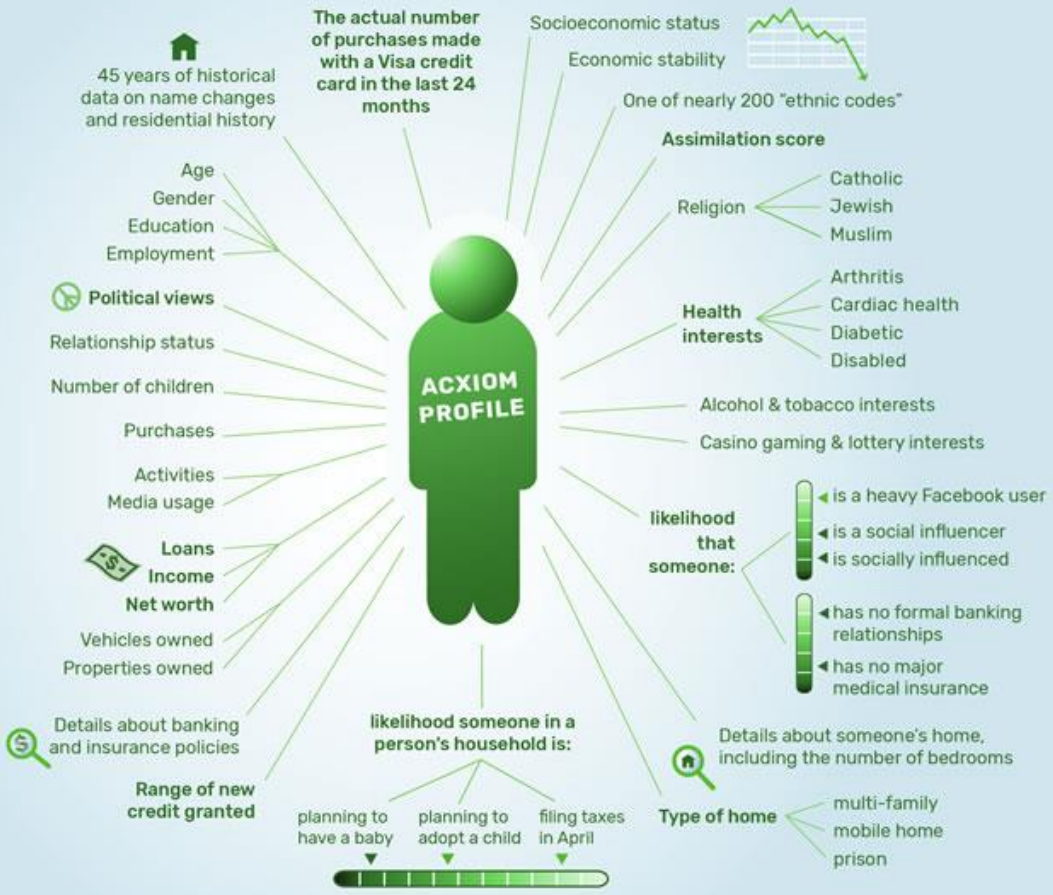
- “It is impossible for me to understand why I am classified and targeted the way I am;
- It is impossible to reconstruct which data any of these segmentations are based on and - most worryingly –
- It is impossible for me to know whether this data can (and is) being used against me.”

(PI investigation)

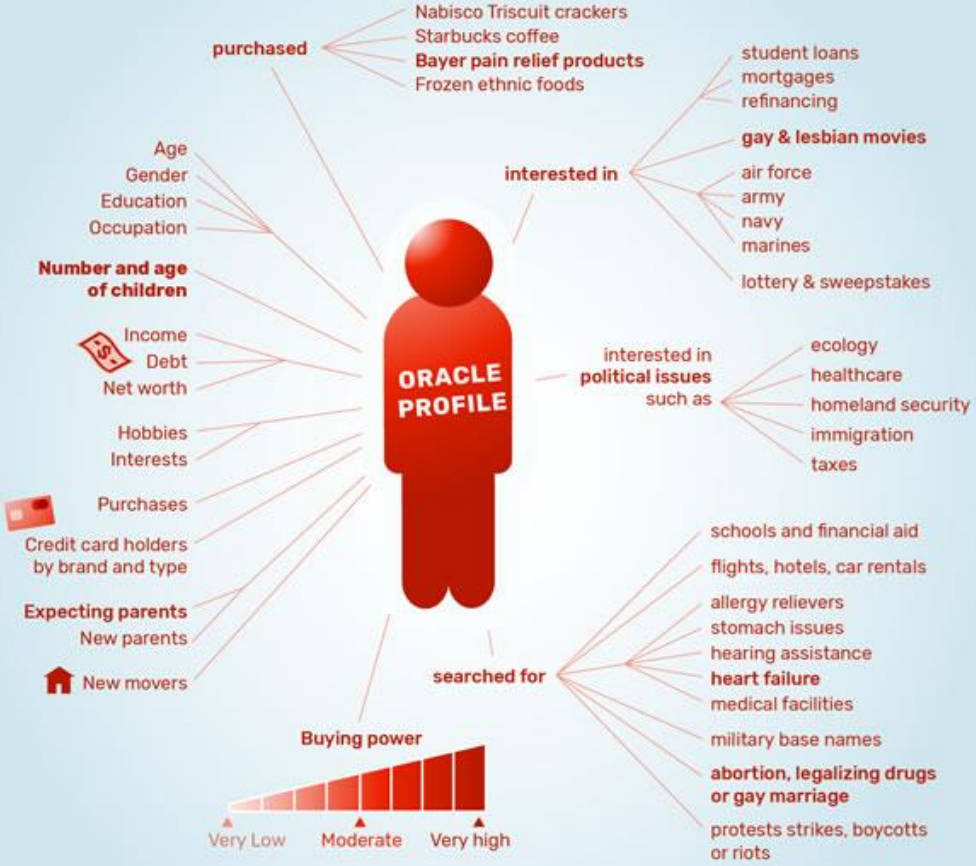
DATA_SEGMENT:Oracle Data Cloud - Oracle UK:Shopping Interests:Hobbies & Interests:Beauty & Style:Jewellery
DATA_SEGMENT:Oracle Data Cloud - MasterCard UK:Shopping Interests:Travel & Leisure:Destinations:Asia
DATA_SEGMENT:Acxiom UK:Shopping Interests:Psychographics & Lifestyles:Lifestyle:Interest in Cultural Pursuits
DATA_SEGMENT:Oracle Data Cloud - MasterCard UK:Shopping Interests:Travel & Leisure:Destinations:Africa
DATA_SEGMENT:Oracle Data Cloud - Oracle UK:Media Interests:TV:Genres:News
DATA_SEGMENT:Acxiom UK:Shopping Interests:Finance:Investment Services:High Investment Activity
DATA_SEGMENT:Oracle Data Cloud - Affinity Answers (UK):Shopping Interests:Social Affinity:Food & Beverage:Cow & Gate
DATA_SEGMENT:Oracle Data Cloud - Oracle UK:Shopping Interests:Consumer Electronics:Brand Affinity:Toshiba
DATA_SEGMENT:Oracle Data Cloud - Experian UK:Shopping Interests:Psychographics & Lifestyles:Mosaic UK:Rural Reality
DATA_SEGMENT:Oracle Data Cloud - Oracle UK:Shopping Interests:Finance:Financial Services:Investing Seekers:College Savings
DATA_SEGMENT:Oracle Data Cloud - MasterCard UK:Shopping Interests:Travel & Leisure:Destinations:Northern Europe
DATA_SEGMENT:Oracle Data Cloud - Oracle UK:Shopping Interests:Hobbies & Interests:Internet & Online Activities:Social Influence
DATA_SEGMENT:Oracle Data Cloud - Oracle UK:Shopping Interests:Automotive:In Market Make:BMW
DATA_SEGMENT:Acxiom UK:Shopping Interests:Psychographics & Lifestyles:Lifestyle:Interest in Gym/Classes
DATA_SEGMENT:Acxiom UK:Shopping Interests:Travel & Leisure:Travel Activities:World Travel as a regular hobby
DATA_SEGMENT:Oracle Data Cloud - MasterCard UK:Shopping Interests:Services:In Market:Hair Care & Beauty Salons
DATA_SEGMENT:Oracle Data Cloud - MasterCard UK:Shopping Interests:Retail:In Market:Jewellery & Giftware
DATA_SEGMENT:Oracle Data Cloud - Datalogix UK:Shopping Interests:Psychographics & Lifestyles:Personas:Avid Readers
DATA_SEGMENT:Oracle Data Cloud - Affinity Answers (UK):Shopping Interests:Social Affinity:Consumer Products:Aveeno

DATA BROKERS HAVE EXTENSIVE PROFILE INFORMATION ON ENTIRE POPULATIONS

Examples of data on consumers provided by Acxiom and Oracle



Acxiom provides of up 3,000 attributes and scores
on 700 million people in the US, Europe, and other regions.



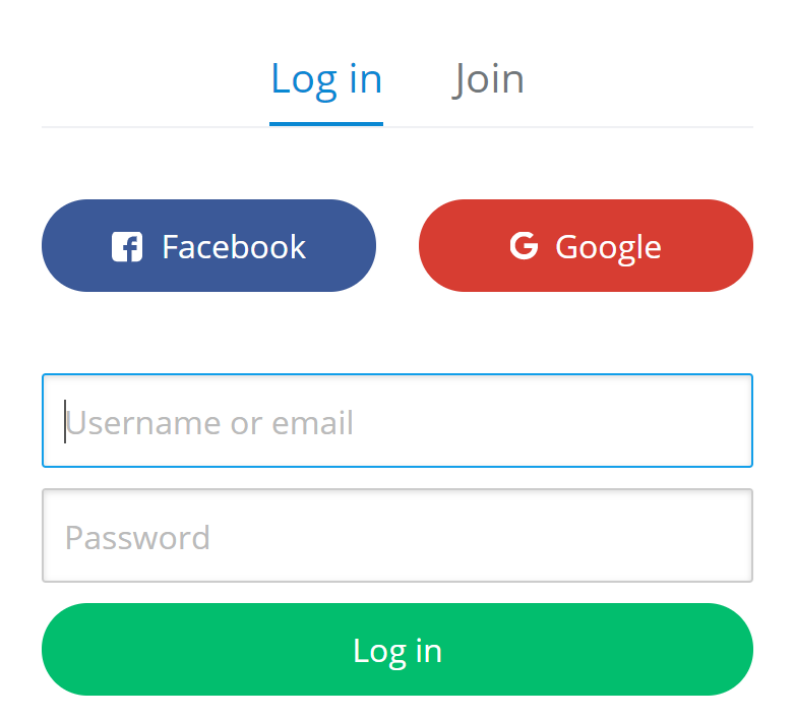
Oracle sorts people into thousands of categories
and provides > 30,000 attributes on 2 billion consumer profiles

© Cracked Labs CC BY-SA 4.0, April/May 2017. Disclaimer: the mentioned companies typically keep information about their activities secret. This illustration is based on publicly available information by Acxiom and Oracle. Every effort has been made to accurately interpret and represent the companies' activities, but we cannot accept any liability in the case of eventual errors. Sources: Acxiom annual reports, developer website (API docs), Oracle press release, help center website, audience playbook, taxonomy updates for January, 2017 (Excel document). For details about the sources see the report "Corporate Surveillance in Everyday Life".

Source: CrackedLabs 2017

Shoshanna Zuboff: Surveillance capitalism

- Data – Prediction products – Behavior modification
  - Business model
  - Scale
  - Consolidation
- Interviews: The Intercept, The Guardian



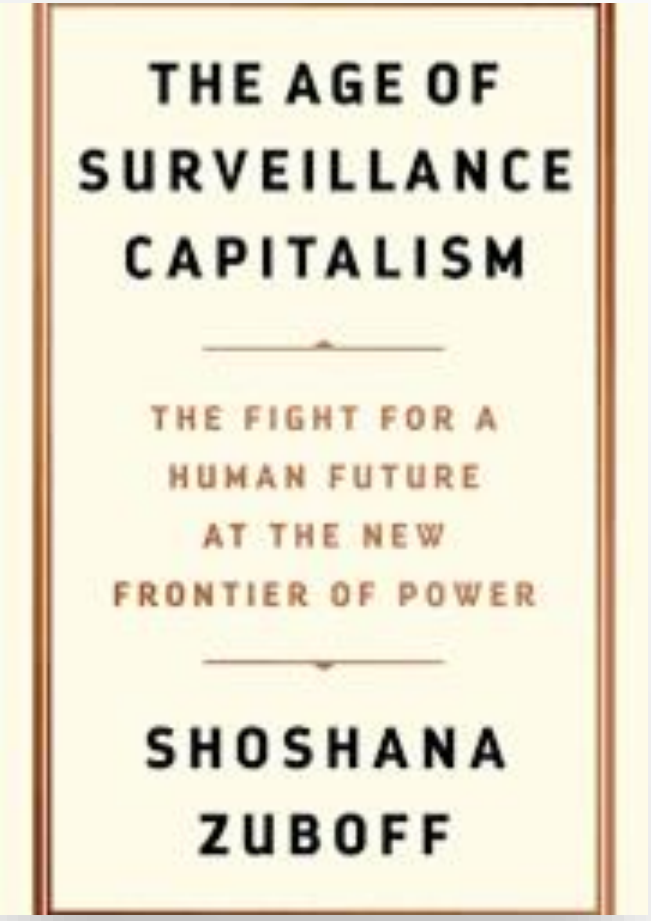
Surveil-lance Cap-i-tal-ism, n.

- **1.** A new economic order that claims human experience as free raw material for hidden commercial practices of extraction, prediction, and sales; **2.** A parasitic economic logic in which the production of goods and services is subordinated to a new global architecture of behavioral modification; **3.** A rogue mutation of capitalism marked by concentrations of wealth, knowledge, and power unprecedented in human history; **4.** The foundational framework of a surveillance economy; **(..) 8.** An expropriation of critical human rights that is best understood as a coup from above: an overthrow of the people's sovereignty.

Shoshanna Zuboff, Surveillance capitalism

Surveillance capitalism

- “Once we searched Google, but now Google searches us. Once we thought of digital services as free, but now surveillance capitalists think of us as free.”
- “Economies of action”: systems designed to modify our behavior for commercial outcomes
- “It is no longer enough to automate information flows about us; the goal now is to automate us.”
- “We are trapped in an involuntary merger of personal necessity and economic extraction”



Concluzii

- Predicția / anticiparea este necesară acțiunii
- Tehnici de analiză a datelor pot identifica patternuri
  - Modelele „prezic” datele, adică reproduc patternuri trecute
  - Oamenii prezic / anticipează viitorul, pornind de la modele și asumpții politice despre viitor
- Tehnici de analiză a datelor
  - Reducerea dimensionalității, analiză cluster, analiză de rețea
  - Analiză de regresie, serii de timp
- Controverse curente
  - Decizii algoritmice și părtinire algoritmică (algorithmic bias)
  - Capitalismul datelor, AdTech și autonomia individuală

SYSTEMS LABORATORY

Bibliografie

PR network crunch

- [1] Wolfie Christl, 2017. Corporate Surveillance in Everyday Life. Cracked Labs.
- [2] John Naughton, 2019, 'The goal is to automate us': welcome to the age of surveillance capitalism
- [3] Norwegian Consumer Council 2020. Out of control. How consumers are exploited by the online advertising industry.
- [4] NYTimes – The Privacy Project
- [5] Frederike Karltheuner, Privacy International, "I asked an online tracking company for all of my data and here's what I found"

SYSTEMS LABORATORY

Cursul DS06

Producția modelelor ML. Explainable AI.

PR network crunch

Structura cursului

- Introducere: De ce DS?
- Scopul cursului
- Măsurarea și colectarea datelor
- Erori de măsurare
  - Precizia
  - Validitatea
  - Fidelitatea
- Corelația
- Erori de modelare
  - Corelații iluzorii
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erori de eșantionare
- Tehnologiile datelor: programare vs. Învățare automată
- Procesul învățării automate
- Machine Learning și Deep Learning
- Predicție prin analiza datelor
  - Analiza factorială (reducerea dimensionalității)
  - Clasificare (cluster)
  - Analiza de rețea
  - Analiza de regresie
  - Predicții și serii de timp
- Predicția în capitalismul datelor

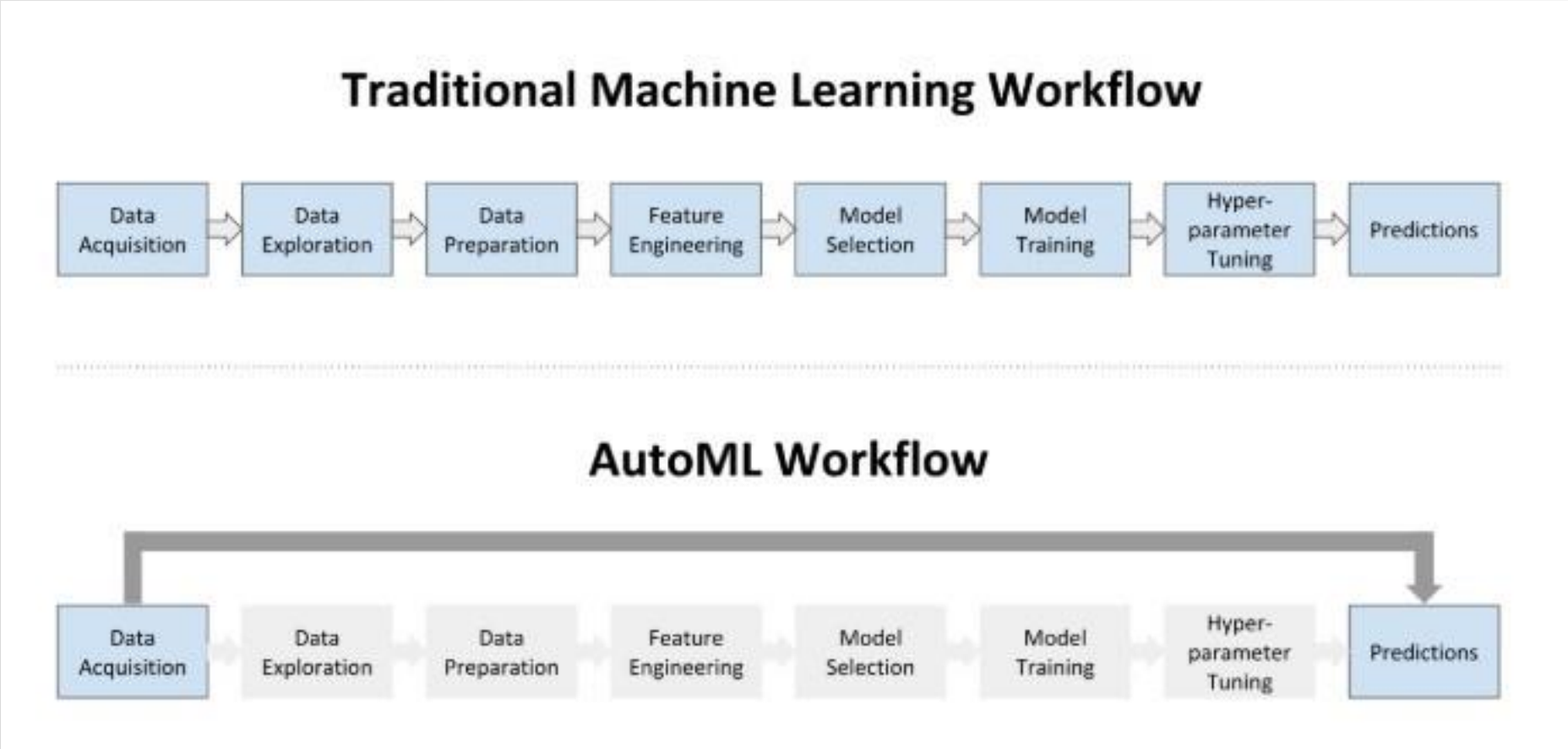
- Auto-ML
- Explainable AI
- ML în lumea reală

#1 Auto-ML

Folosirea ML pentru a construi ML

Auto-ML

- Automatizarea dezvoltării, evaluării și rafinării modelelor



Auto-ML: Input - Output

- Input: setul de date și problema de rezolvat
- Output: rafinarea datelor, pregătirea și specificarea modelului, predicții
- Similar cu identificarea celei mai bune arhitecturi pentru piesele LEGO disponibile

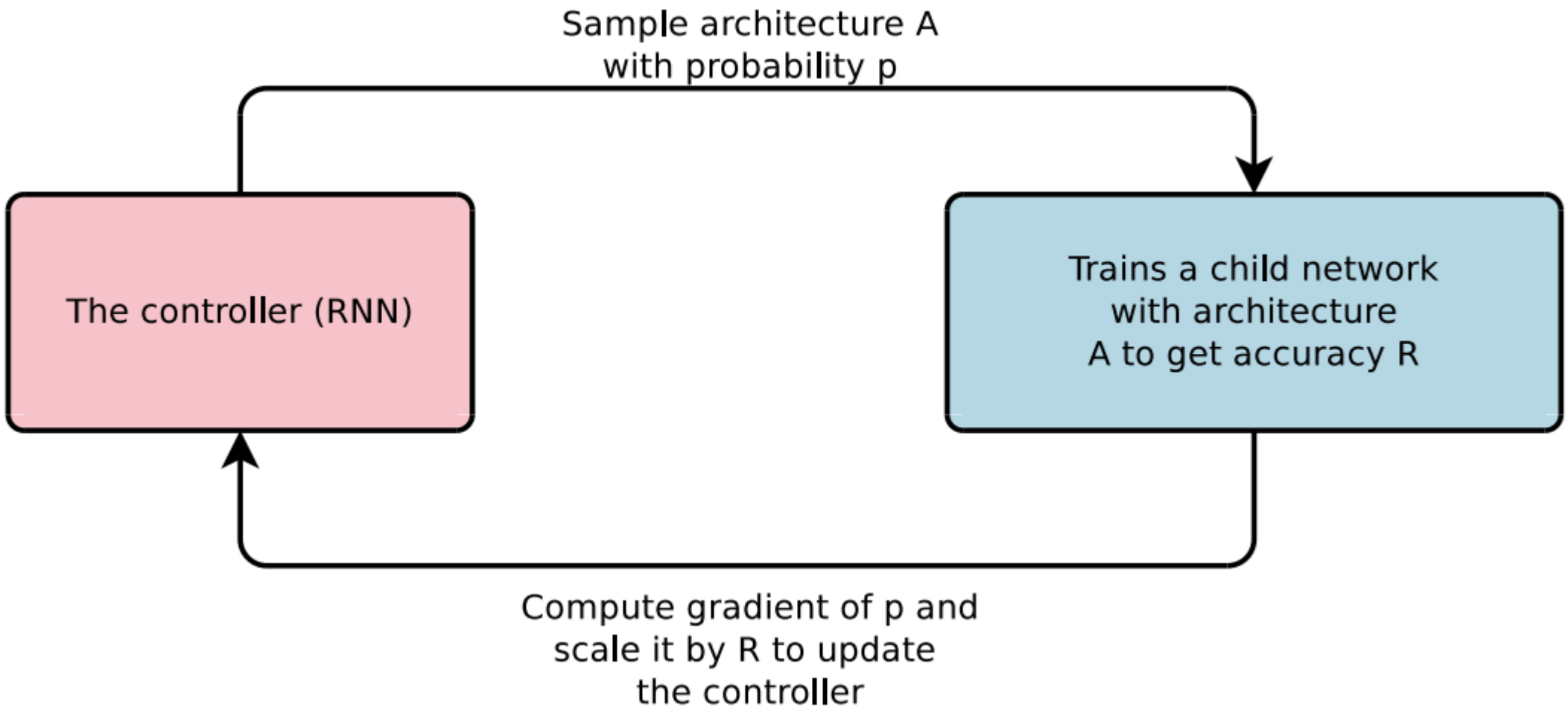
Auto-ML: Dezavantaje

- Este **extrem** de costisitor computațional. Numai giganții industriei își pot permite o implementare, și numai ei oferă soluții.
- Implică **folosirea API-ului** lor pentru predicție
- Nu îți poți vedea modelul creat!
  - Nu există privacy, totul trecând prin cloud
  - Nu poți da deploy de un device offline sau mic, cum ar fi Raspberry Pi
  - Nu poți învăța ce l-a făcut pe model bun, și asta nu îți amortizează din costul de producție pe termen lung

Auto-ML: Limitări

- Cineva tot trebuie să ofere acele 'bucăți de LEGO' inițiale modelului, altfel căutarea ar fi incredibil de mult mai lungă. Iar alegerea lor nu este deloc trivială.
- Formularea problemei încă rămâne în mâinile noastre, și aceasta de multe ori este cel mai important pas.

Funcționarea auto-ML



#2 Explainable AI

Putem înțelege modelele AI?
Putem înțelege erorile AI?

Revoluția AI

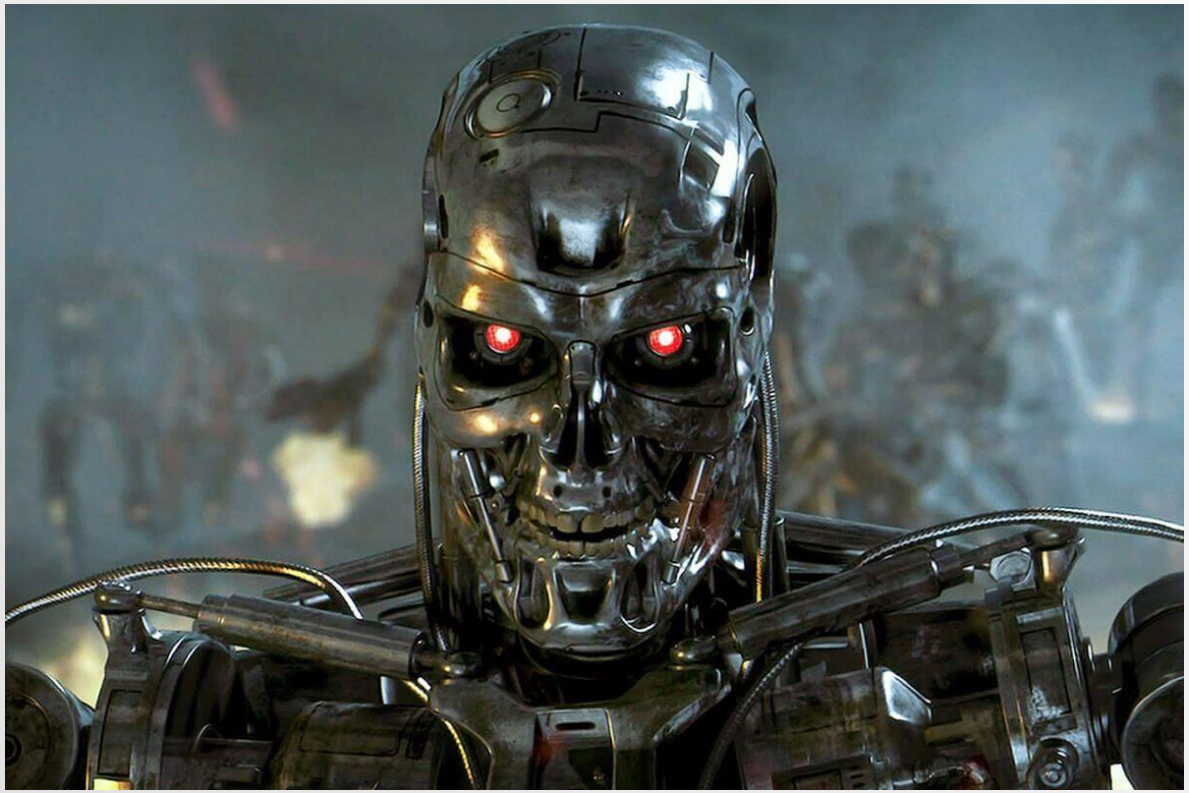
- “Artificial intelligence will be more profound for humanity than fire and electricity.” (Sundar Pichai, CEO Google)
- Focul și electricitatea pot scăpa de sub control! Dar AI?

Controlul uman asupra modelelor AI

- Pe măsură ce complexitatea crește și AI îndeplinește funcții critice…
  - Putem înțelege ce face?
  - Putem înțelege dacă și unde greșește?
- Exemple
  - Discriminare algoritmică
  - Utilizarea AI pentru arme inteligente
  - Atacuri (ex. 1-pixel attack) (Su et al., 2017 – pe scurt)

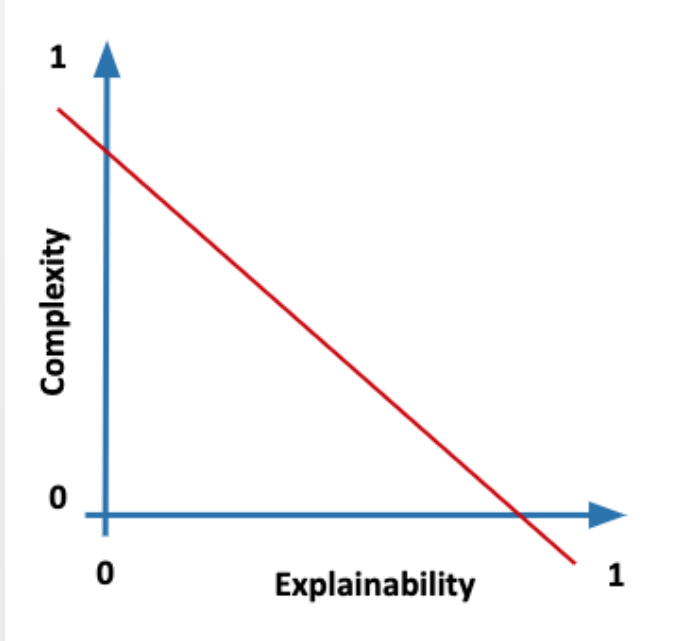
Explainable AI: provocarea | 1

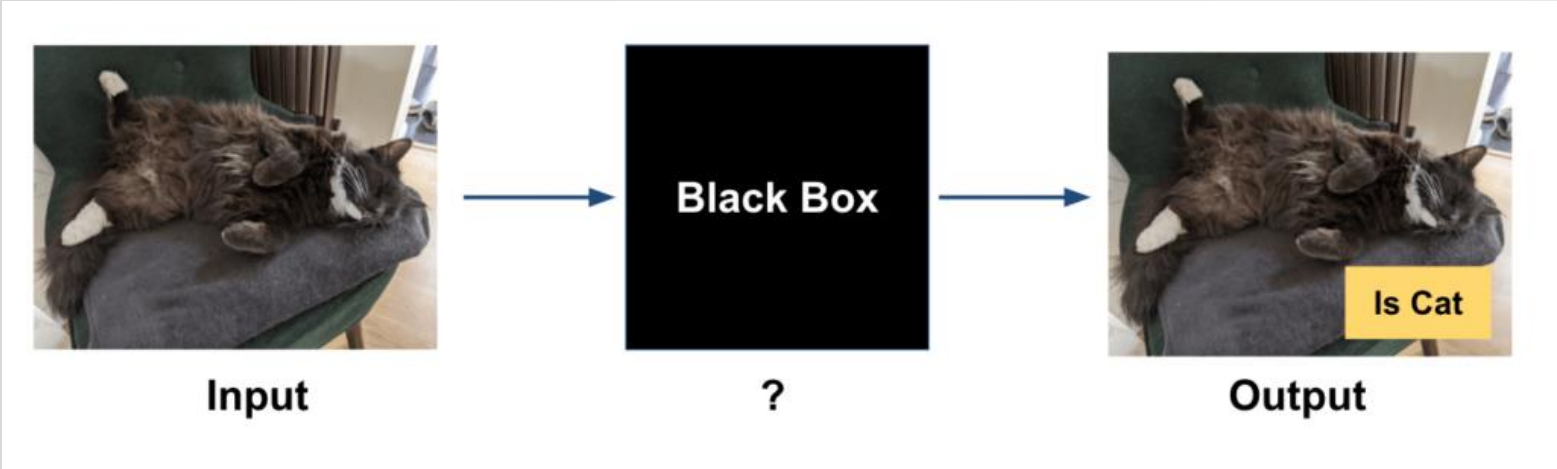
- **Scopul** AI: să rezolve probleme prea complexe pentru abordări algoritmice
- Scopul AI: să scurtcircuiteze înțelegerea umană!



Explainable AI: provocarea | 2

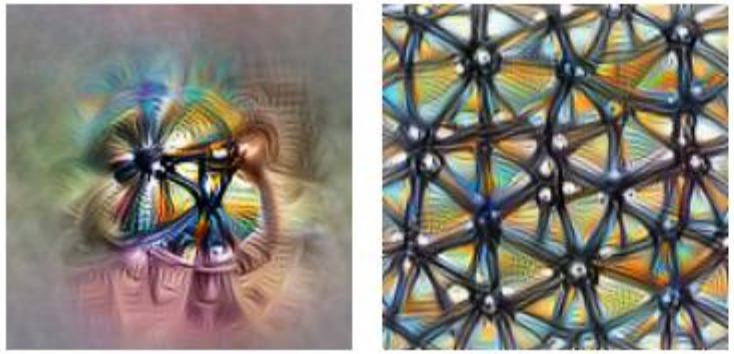
- Complexitatea unui model este:
  - Invers proporțională cu abilitatea noastră de a-l înțelege
  - Direct proporțională cu **performanța**





Vizualizarea RNC | 1

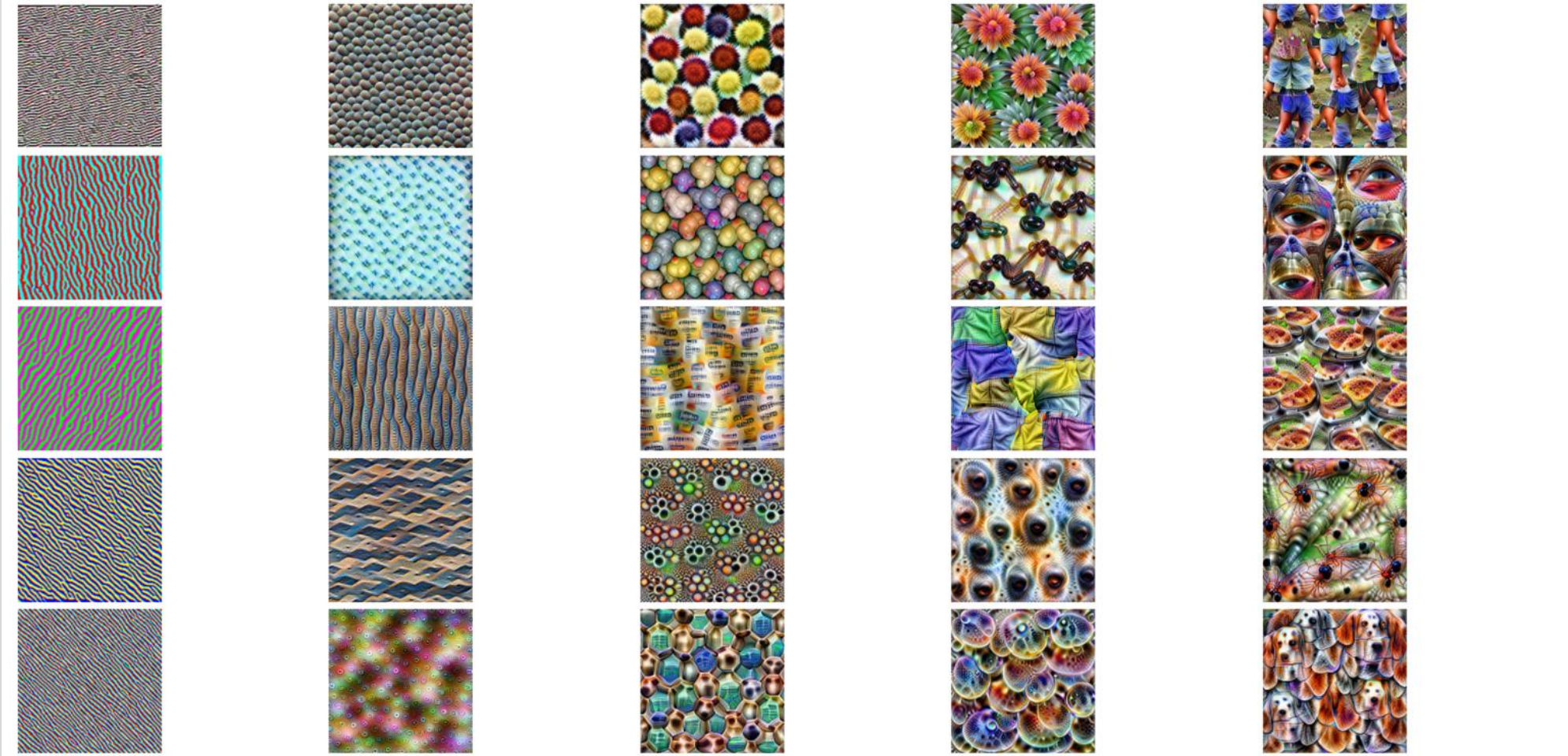
- Prin vizualizare putem înțelege mai bine cum rețelele neurale convoluționale „înțeleg” imaginile (Olah et al., 2017)
  - Vizualizarea trăsăturilor (*feature vizualization*): la ce se uită rețeaua?
  - Atribuirea: ce parte din imagine influențează decizia?



Feature visualization answers questions about what a network — or parts of a network — are looking for by generating examples.

Attribution [1] studies what part of an example is responsible for the network activating a particular way.

Feature vizualization: straturi



Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)

(Olah et al. 2017)

Feature vizualization: ce vede rețeaua?

Dataset Examples show us what neurons respond to in practice

Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.

Baseball—or stripes? *mixed4a, Unit 6*

Animal faces—or snouts? *mixed4a, Unit 240*

Clouds—or fluffiness? *mixed4a, Unit 453*

Buildings—or sky? *mixed4a, Unit 492*

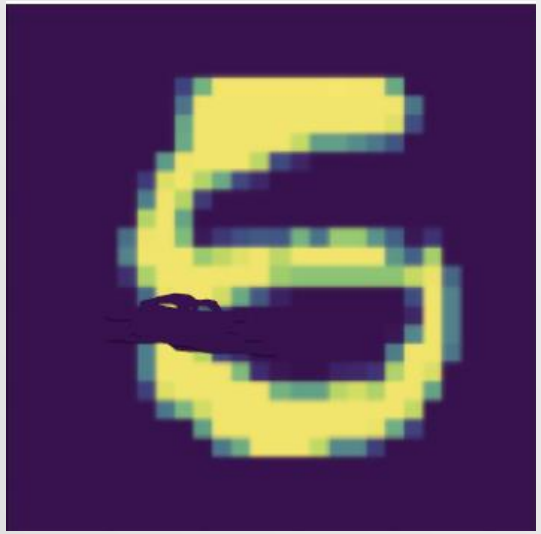
(Olah et al. 2017)

Oameni vs. modele: black box

- Și oamenii și modelele ML sunt cutii negre!
  - Cum decidem ce vedem în imaginea de mai jos? [1]
  - Ce anume ne-ar face să fim mai siguri ca este un 5? [2]

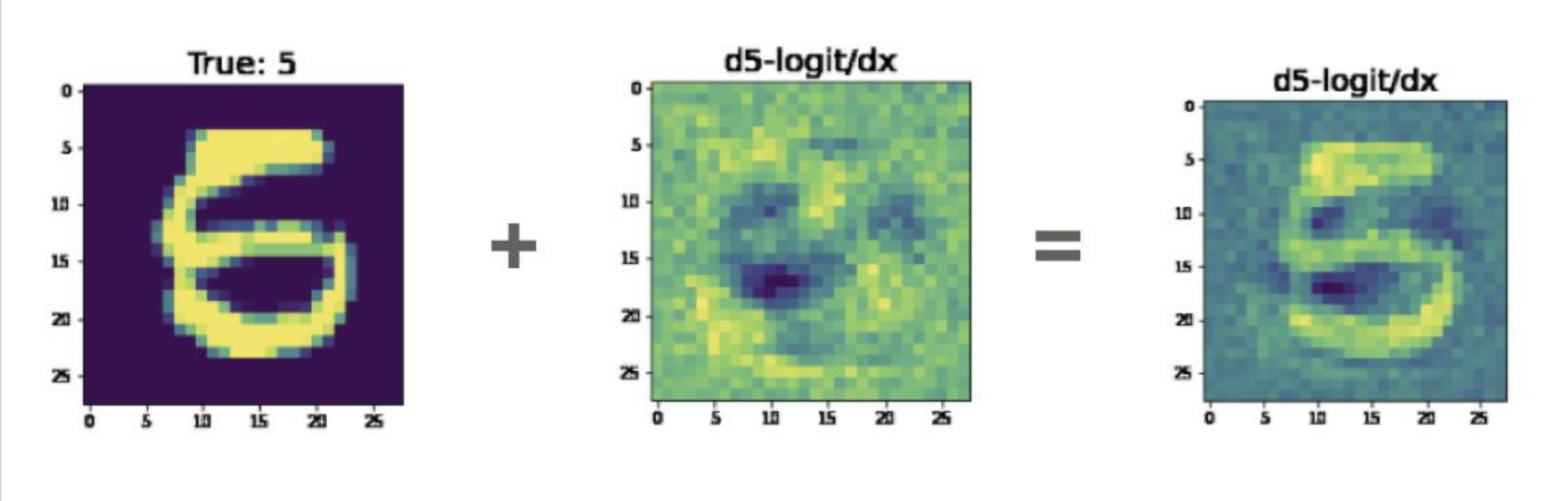
[1]

[2]



Modele simple vs. complexe

- Un model simplu este adesea inteligibil și similar cogniției umane
- Modelele complexe sunt mai greu de înțeles

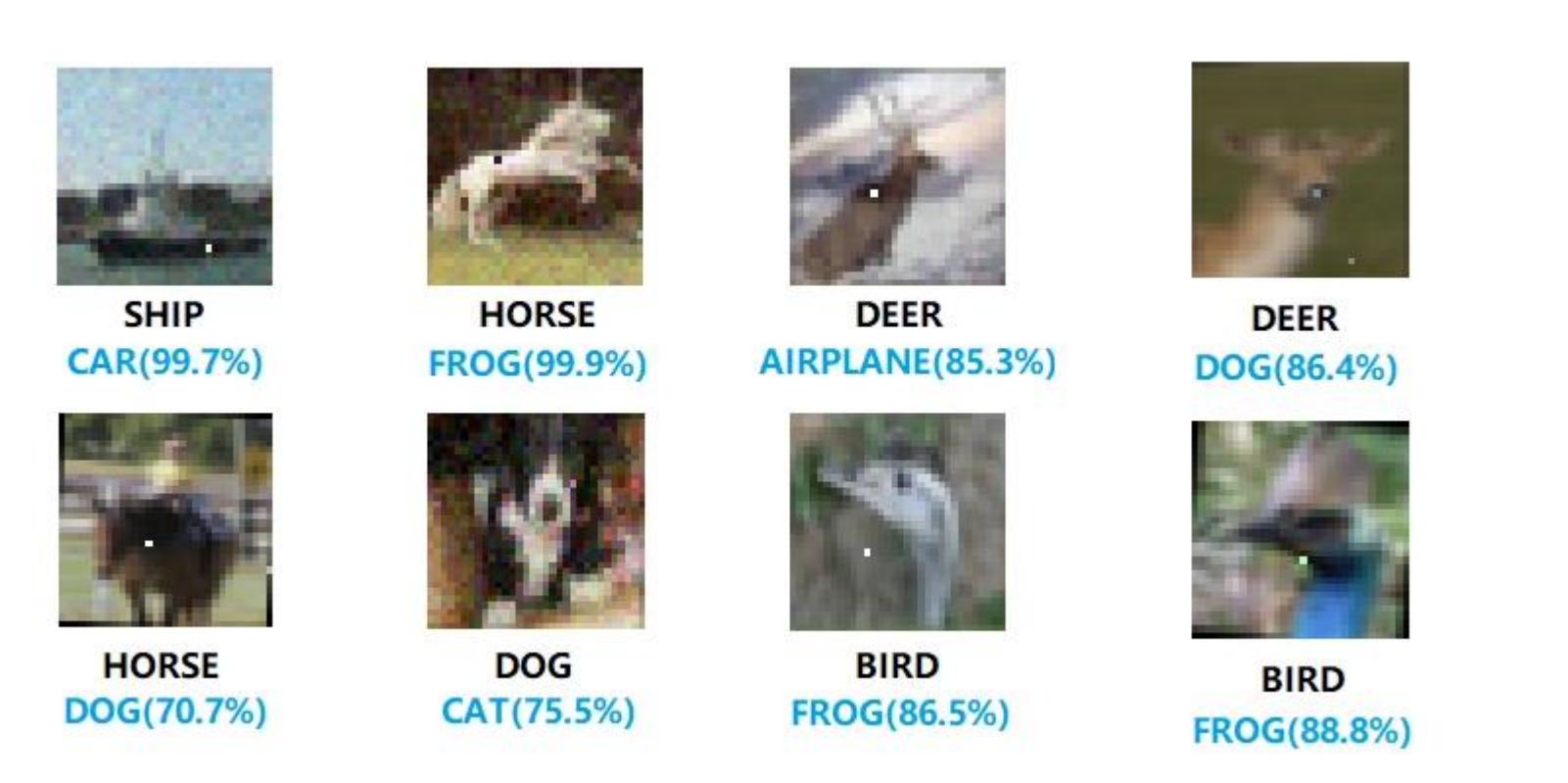


Probleme: algorithmic bias

- August 2020: Apple card oferă femeilor limite de credit mai mici decât bărbaților – chiar și în cupluri cu venituri/conturi comune (Knight 2020)
- De ce? Părtinire algoritmică implicită în date:
  - Apple: „No one from the company seemed able to describe how the algorithm even worked, let alone justify its output”
  - Goldman-Sachs: „it doesn’t even use gender as an input”

Probleme: 1-pixel attack

- Prin schimbarea unui pixel bine ales modelul reinterpretează radical imaginea (Su et al., 2017 – pe scurt)



Tesla attack

- „Tencent's Keen Security Lab showed how they were able to bamboozle a Tesla Model S into switching lanes so that it drives directly into oncoming traffic." (Tencent 2019) – pe scurt
  - Prin plasarea unor stickere ce creează „iluzia" că drumul coteşte

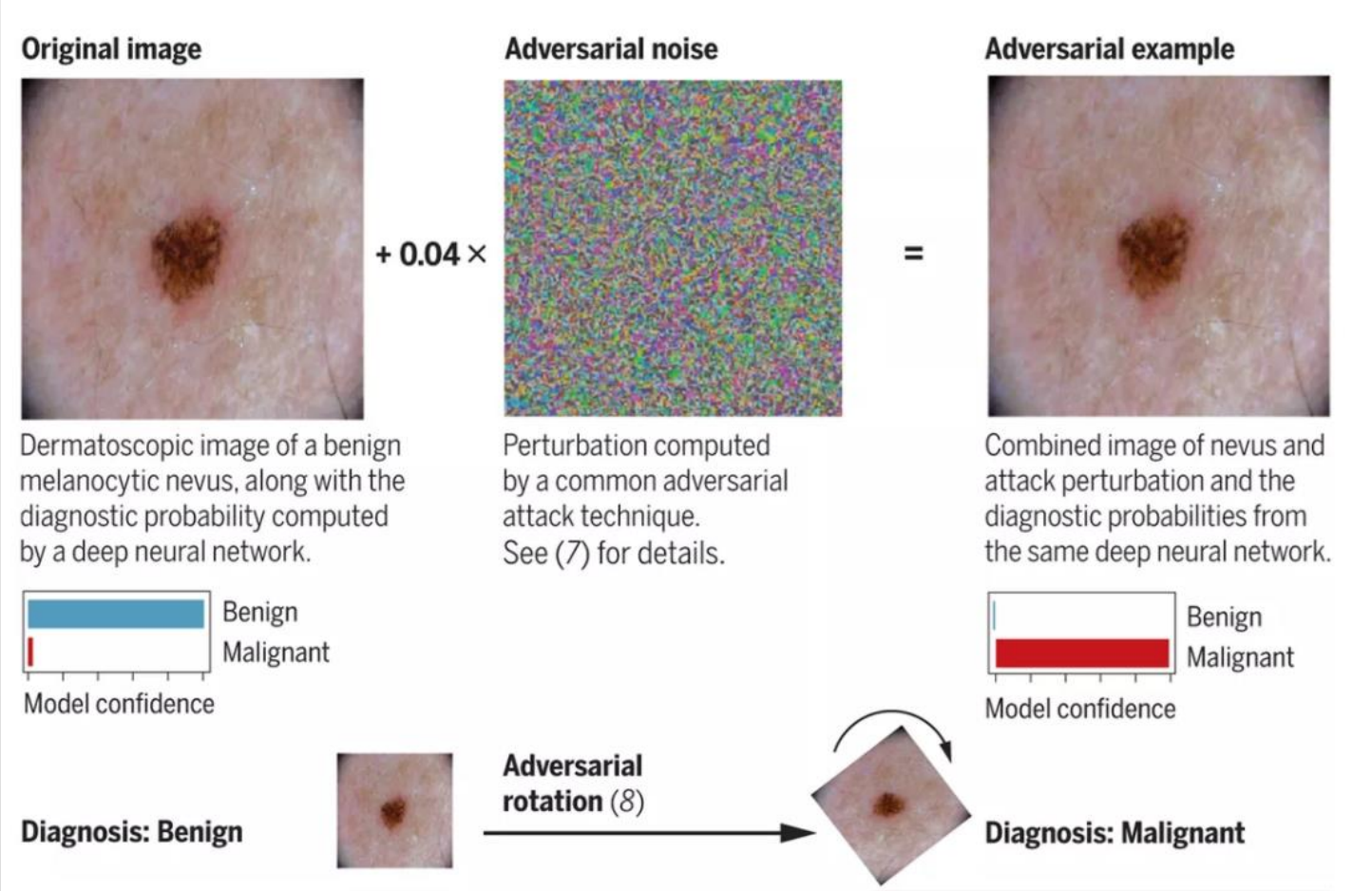
Fig 34. Fake lane mode in physical world

Fig 35. In-car perspective when testing, the red circle marks, the interference markings are marked with red circles

Alte posibile scenarii de atac

- Generarea unor mesaje email care să inducă generatoarele de mesaje automate **să divulge numărul cardului de credit**
- Păcălirea diagnozelor AI pentru a justifica **intervenții medicale mai scumpe**

(Samuel 2019)



#3 ML în lumea reală

PR network crunch

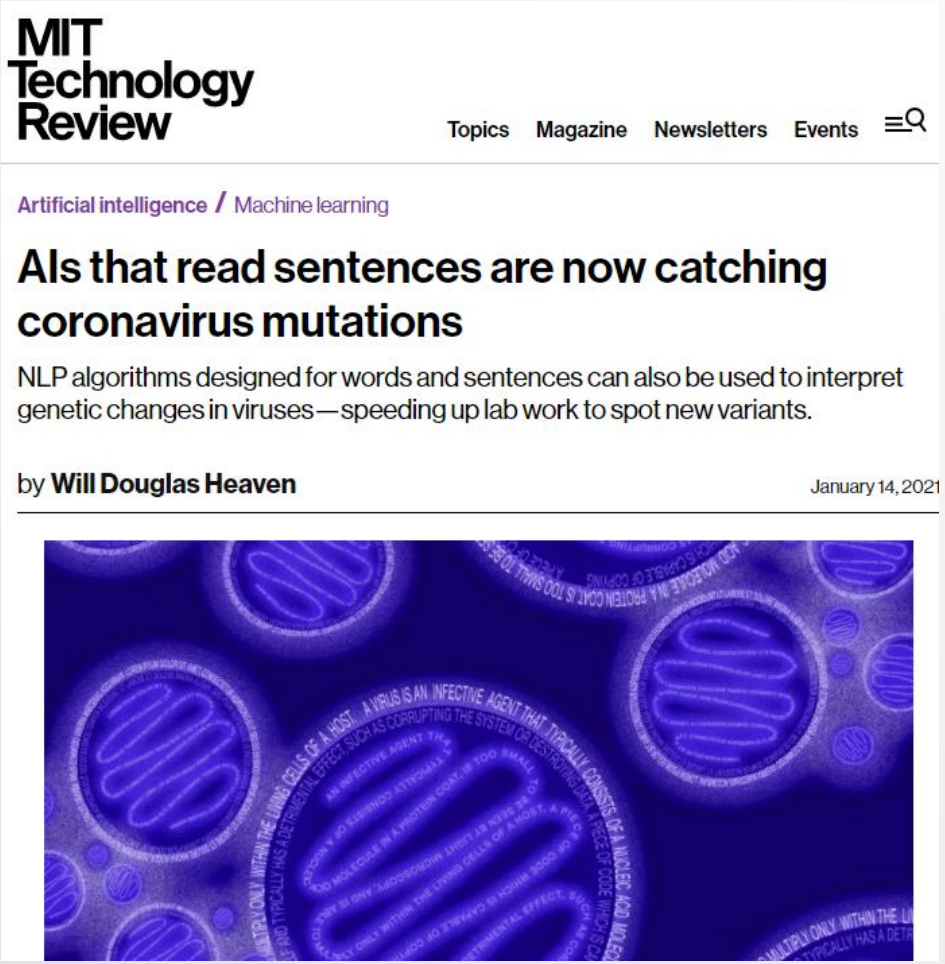
Procesul ML

- Proces iterativ
- Pregătire și antrenare pe date „de laborator”
- Producție: intrarea în lumea reală

1. Colectare
2. Pregătire
3. Antrenare
4. Evaluare
5. Producție
6. Re-antrenare

Real-life AI: Viruși = limbaj?

- Modele LSTM pot fi antrenate să detecteze posibilele mutații virale ce mențin/sporesc rata de infecțiozitate
- Sistemul imunitar interpretează virusurile ca pe un limbaj (Heaven, 2021)
  - „The overall aim of the approach is to identify mutations that might let a virus escape an immune system without making it less infectious—that is, mutations that change a virus's meaning without making it grammatically incorrect"
- De ce LSTM? Scalare:
  - „These older networks can be trained on far less data than transformers and still perform well for many applications"



MIT Technology Review

Topics Magazine Newsletters Events

Artificial intelligence / Machine learning

AIs that read sentences are now catching coronavirus mutations

NLP algorithms designed for words and sentences can also be used to interpret genetic changes in viruses—speeding up lab work to spot new variants.

by **Will Douglas Heaven** January 14, 2021

Scalarea sistemului

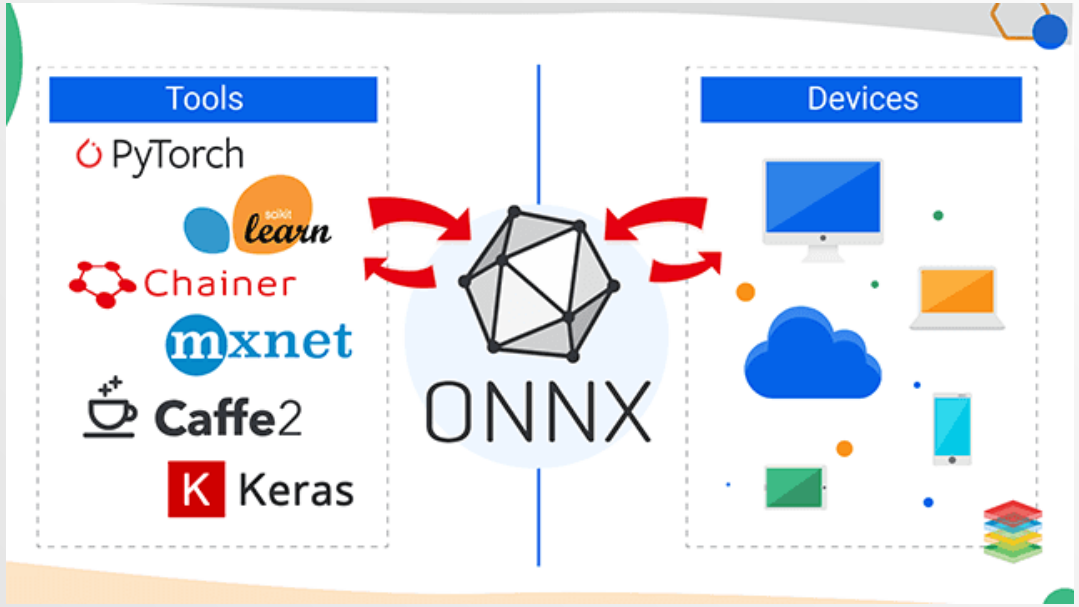
- Scalarea modelelor ML e foarte costisitoare
- Două opțiuni pentru a-mi oferi modelul ca un serviciu:
  - **Client-based**: Pe mașina clientului, fiecare utilizator rulând local o versiune a modelului meu pentru a primi răspuns la query.
  - **Server-based**: Pe un server/cluster de servere, astfel datele vor ajunge la mine, și le pot rula pe hardware specific și performant.

Client-based: Beneficii

- Nu trebuie să construiesc o arhitectură complicată și scalabilă pe orizontal (să nu mai spunem scumpă!)
- Mentenanța este mică.
- Nu există latență de comunicare cu serverul.
- Nu avem de a face cu data privacy, totul fiind local.

Client-based: Provocări

- Utilizatorii pot rula în **medii diverse**
  - Diverse sisteme de operare, diverse browsere, diverse dispozitive
- Soluții pentru compatibilitate:
  - **tensorflow.js**: rularea modelelor în browser, indiferent de backend
  - **ONNX** (Open Neural Network Exchange) suportă un număr mare de platforme hardware, de la cele mai optimizate componente de procesare vectorială (AVX512), până la Raspberry Pi

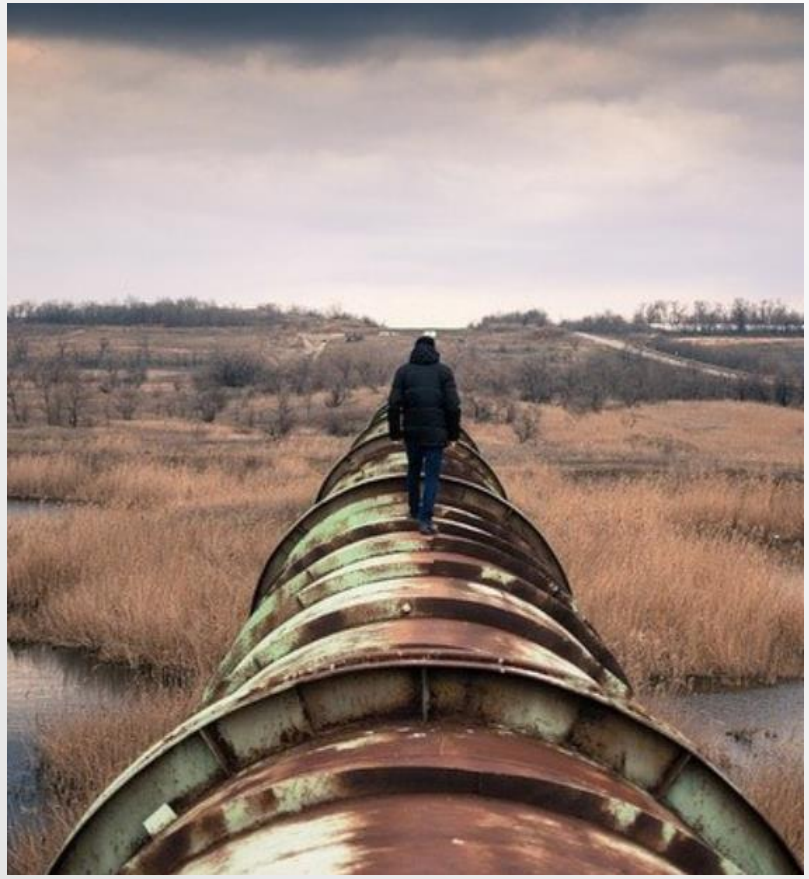


Client-based: Probleme

- Performanță nu este asigurată
  - Dacă dorim să asigurăm un anumit grad de performanță nu putem să lăsăm toată procesarea pe client.
- Chiar modelele intens optimizate sunt scumpe și pot fi o provocare pentru anumite dispozitive.
- Un număr mare de probleme nu se pretează la un astfel de model, datele procesate fiind de interes pentru noi!
- Modelele tind să aibă nevoie de multă pre/post procesare și feature-engineering
  - Chiar dacă modele în sine pot fi portate, algoritmii externi de procesare nu pot, și de multe ori trebuie rescriși

Server-based: Provocări

- Pentru a ne scala modelele, avem nevoie de sisteme ce ne permit distribuirea și procesarea eficientă a datelor peste un cluster de calculatoare.
- Totul trebuie făcut în paralel
  - Încărcare, procurare, procesare, împărțire și stocare
- Sistemul trebuie să fie performant, valabil și tolerant la defecte!
- **E necesar un pipeline** compus din soluții dedicate
  - Procesăm rapid (scalabil orizontal) cantități imense de date,
  - chiar dacă folosim algoritmi costisitori (precum modelele de DL)



Server-based: Soluții | 1

- **Hadoop MapReduce:** procesare și generate de cantități mari de date prin algoritmi paraleli și distribuiți pe un cluster.
- **HDFS(Hadoop Distributed File System):** un sistem de fișiere distribuit, cu blocuri de date de dimensiuni mari și replicare implicită.
- **Spark:** interfețe pentru programarea pe clustere cu paralelism de date și rezistență la defecte implicită; foarte rapid
- **Mesos:** un cluster manager ce organizează resursele de procesare la un nivel foarte granular, sporind colaborarea nodurilor.

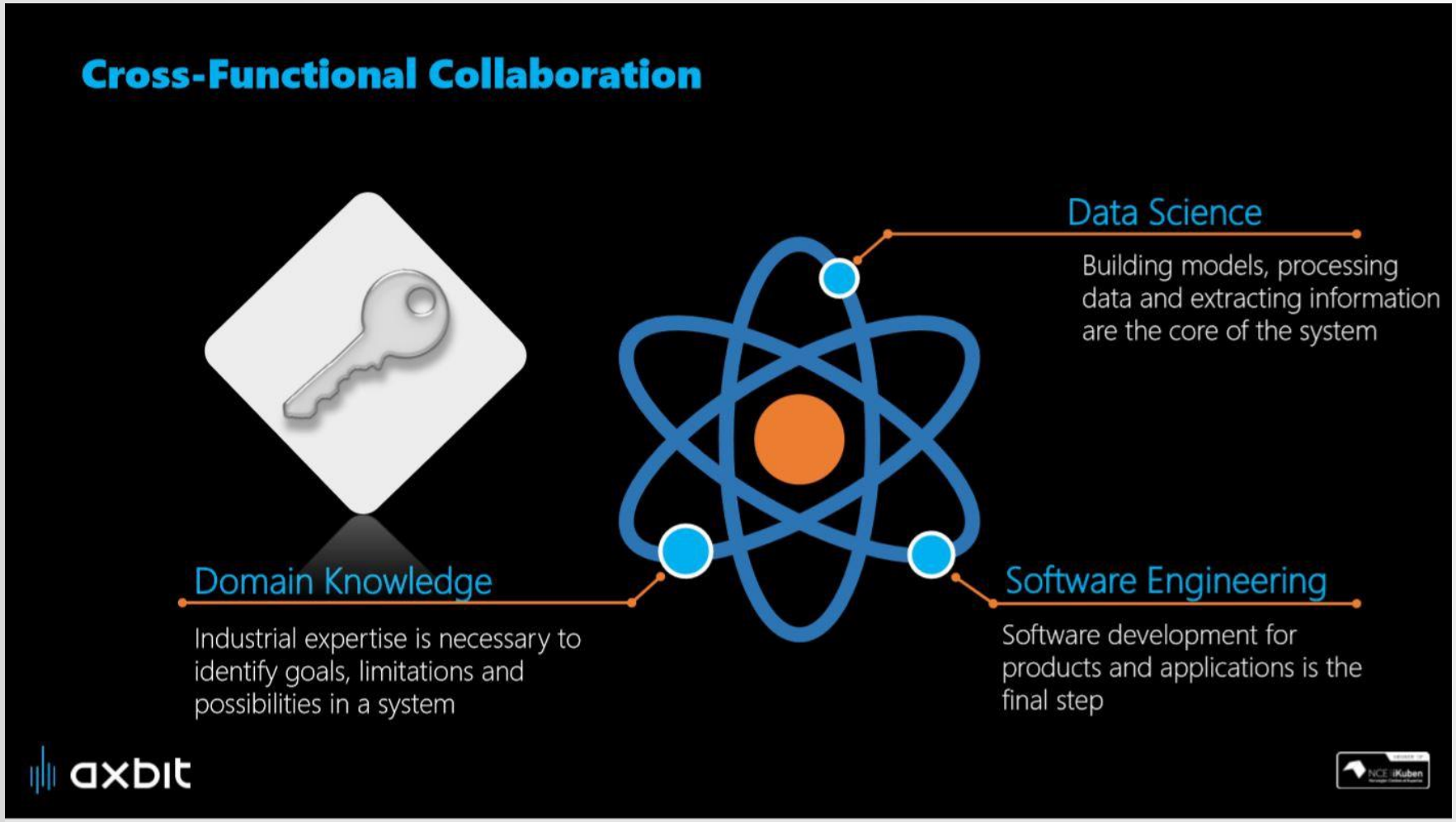
Server-based: Soluții | 2

- **ZooKeper:** sistem distribuit pentru păstrarea ierarhică de perechi key-value, ce oferă sincronizare și registre de nume pentru sisteme distribuite.
- **Cassandra:** o bază de date NoSQL proiectată să lucreze cu cantități mari de date în mod distribuit, oferind o valabilitate înaltă și redundanță.
- **Avro:** un format de date care permite serializarea într-o formă binară foarte compactă, și rezilientă, care poate fi trimisă prin HTML.
- **Kafka:** o platformă de stream processing capabilă să ofere high throughput și latență scăzută, foarte util pentru lucrul cu date în timp real.

Server-based: Probleme

- **Costurile** sunt ridicate
  - Implementare & întreținere
- **Eficiența** implementării
- **Latența**
  - Poate fi o problemă dacă dorim răspunsuri în timp real
- **Privacy**
  - Criptarea datelor

Server-based: Alte ingrediente



Server-based: alte ingrediente

- **Cunoștințe de domeniu**
  - Avem nevoie de cunoștințe din lumea problemei pe care încercăm să o rezolvăm
- **Știința datelor**
  - Necesară pentru a extrage tipare din date și a le organiza
- **Ingineria software**:
  - Crearea unei infrastructuri robuste pentru procesarea eficientă și sigură de date
  - Crearea de pipeline-uri și mentenanța lor
  - Livrarea soluțiilor într-un mod prietenos

Referințe

1. Su, J., Vargas, D.V. and Sakurai, K., 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), pp.828-841.
2. Olah, C., Mordvintsev, A., Schubert, L., 2017. Feature Visualization. How neural networks build up their understanding of images. Online.
3. Google Machine Learning Crash Course
4. Knight, W., 2020. The Apple Card Didn't 'See' Gender—and That's the Problem. Wired.
5. Tencent Keen Security Lab, 2019. Experimental Security Research of Tesla Autopilot. Online.
6. Samuel, S. 2019. It’s disturbingly easy to trick AI into doing something deadly. Vox.
7. Heaven, W.D., 2021. AIs that read sentences are now catching coronavirus mutations. MIT Technology Review