Curs 03 – Măsurare

Măsurare și colectare

Distribuții

Modele de măsurare

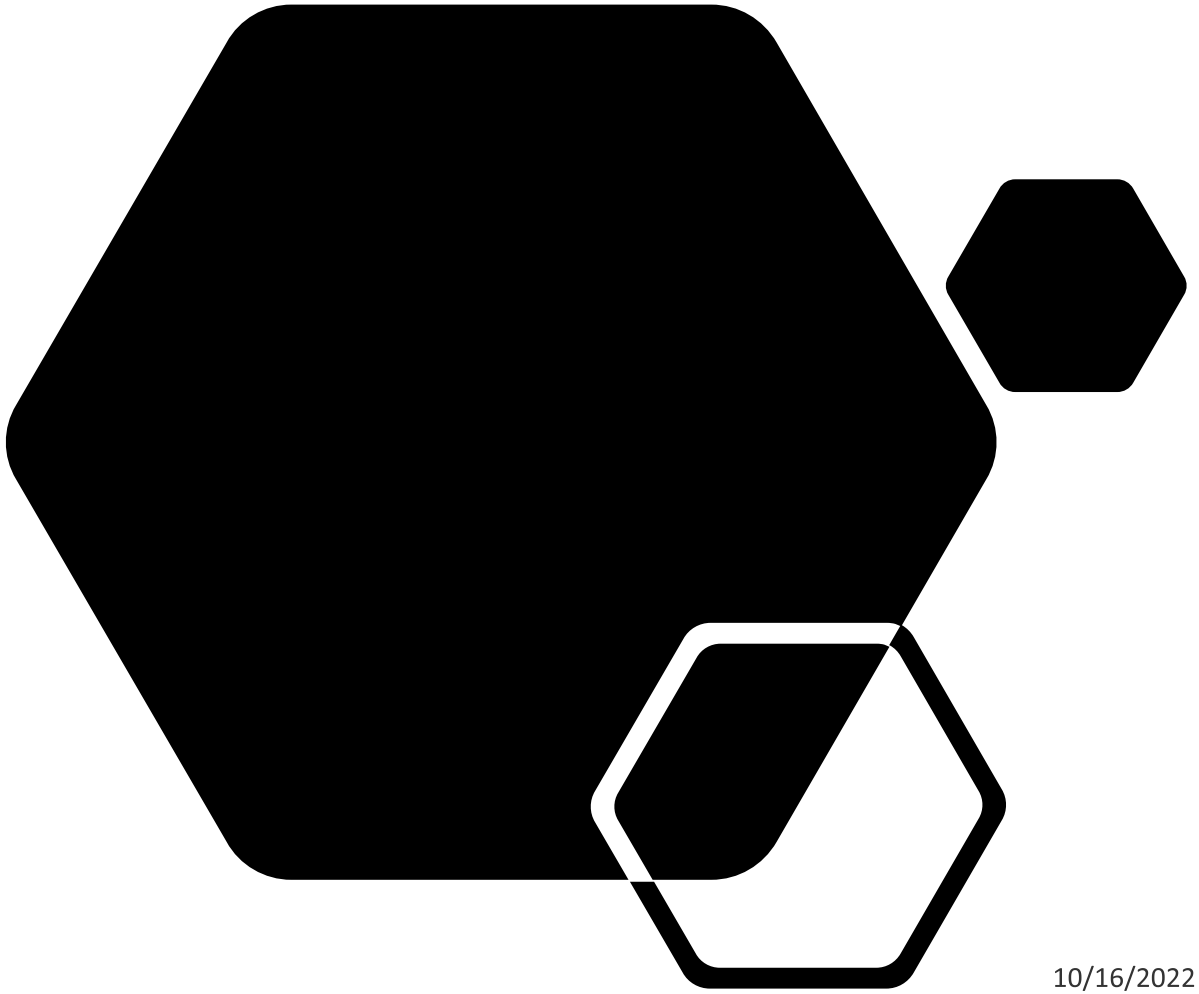
Erori de măsurare

10/16/2022

Structura cursului

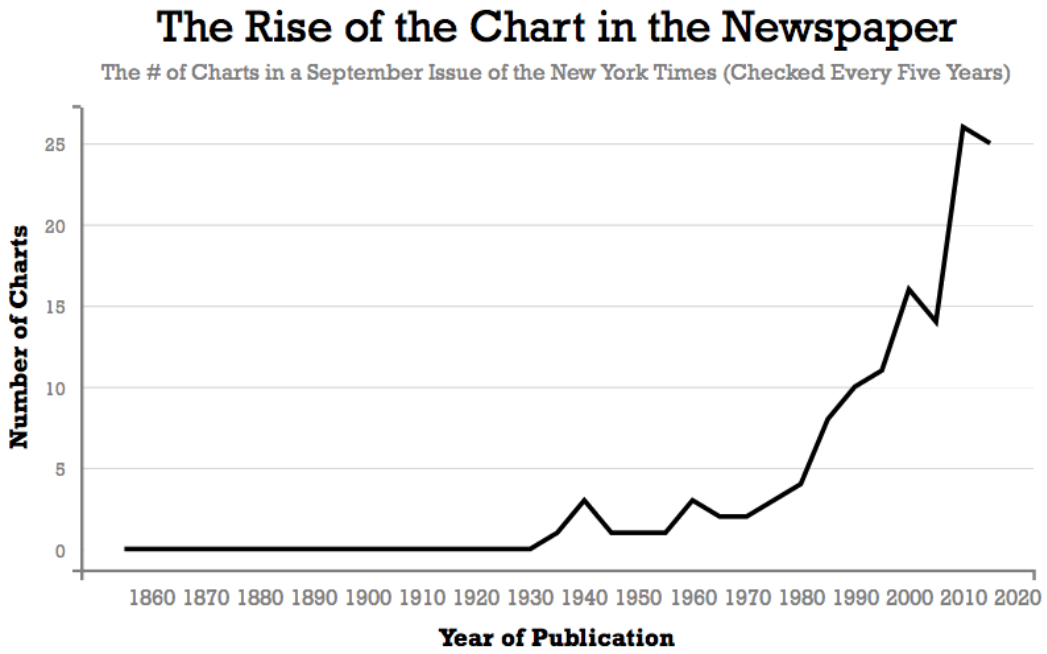
1. Why?
2. Cauzalitate
3. Măsurare
   - Măsurare și colectare
   - Distribuții
   - Modele de măsurare
   - Erori de măsurare
4. Modelare și eșantionare
5. Tehnici de analiză
6. Predicție
7. Programare și ML
8. ML și Deep Learning
9. Producția
10. Why Privacy?
11. Privacy Preserving Algorithms
12. Privacy Architectures and Federated Learning

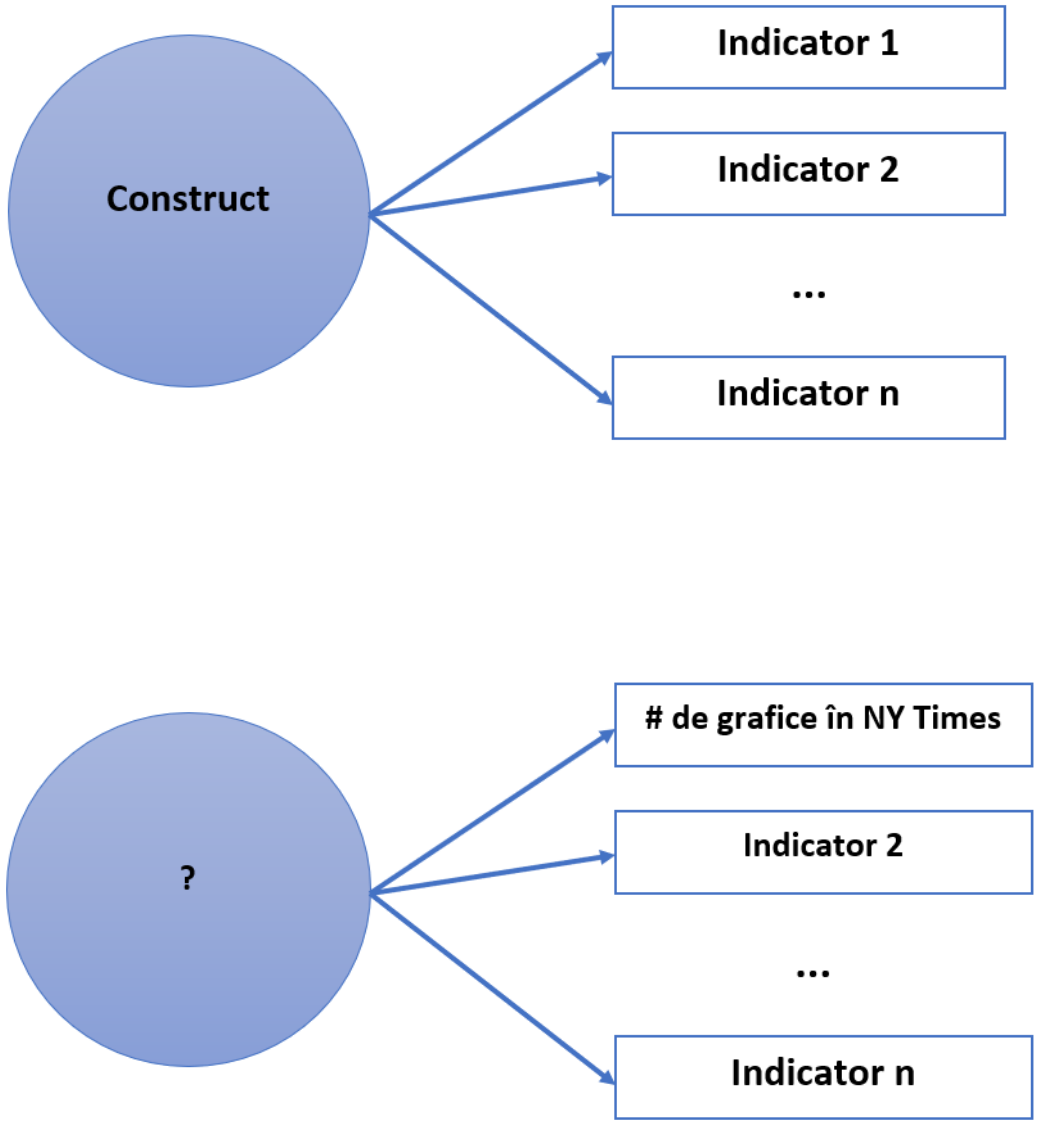
Măsurarea și colectarea datelor



Măsurare vs. colectarea datelor

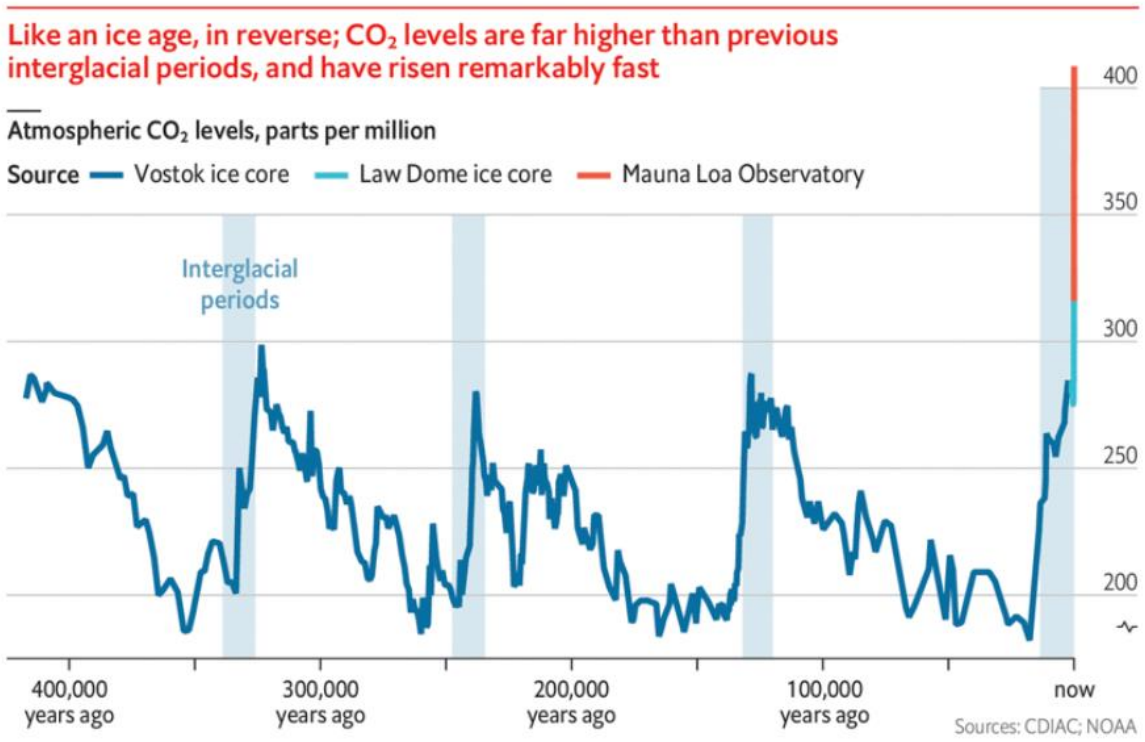
- Măsurarea: transformarea **fenomenelor** în **variabile**
  - Înălțimea → un număr de cm
  - Greutatea → un număr de kg
  - Temperatura → un număr de grade
- Colectarea datelor: organizarea & stocarea datelor din:
  - Măsurători proprii
  - Măsurători din surse diferite
- Heterogenitatea datelor: fiecare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


The Rise of the Chart in the Newspaper
The # of Charts in a September Issue of the New York Times (Checked Every Five Years)
Number of Charts
25
20
15
10
5
0
1860 1870 1880 1890 1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020
Year of Publication


Construct
Indicator 1
Indicator 2
...
Indicator n
?
de grafice în NY Times
Indicator 2
...
Indicator n

Sursa

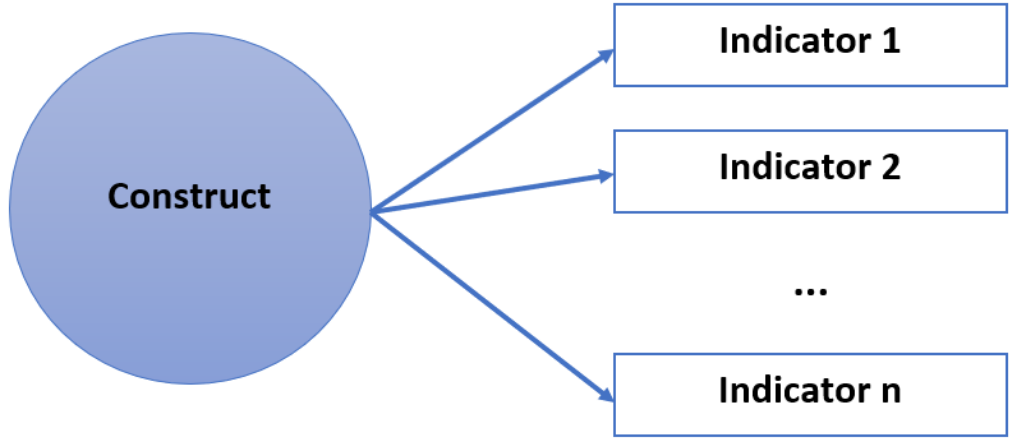
Măsurarea trecutului prin prezent



Măsurarea CO2 cu mii de ani în urmă prin CO2 actual în ghețari

- CO2 în trecut = constructul
- CO2 în prezent, în ghețari = indicatorul

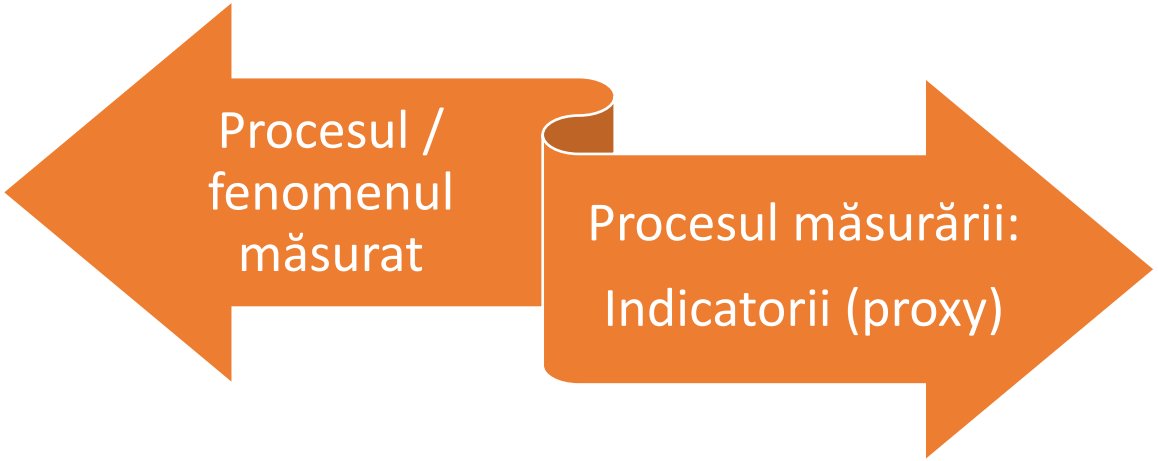
+ Armonizarea și agregarea datelor din surse diferite



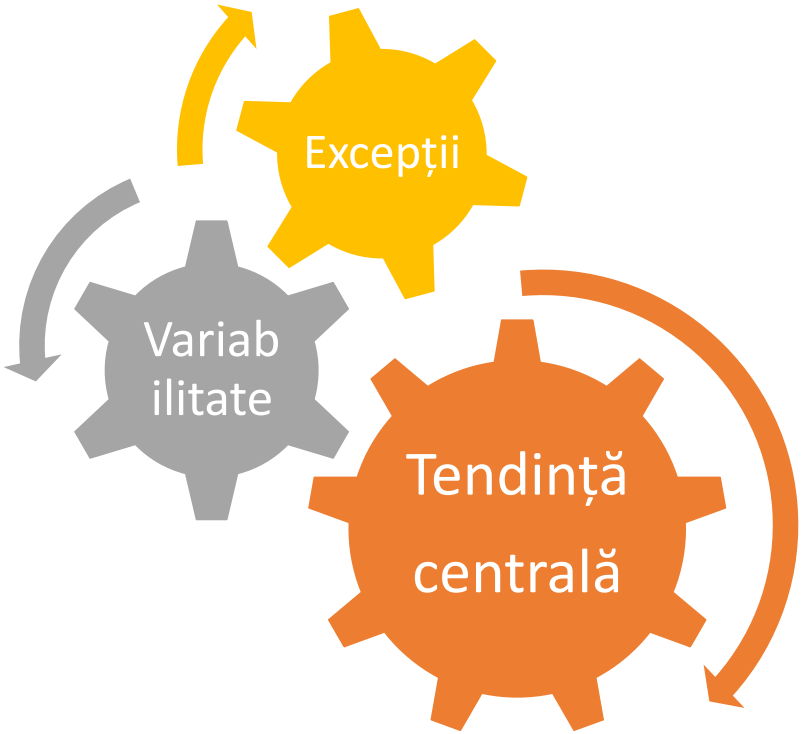
Source:
The Economist

Două competențe centrale în DS

Să facem diferența între:



Să ținem cont mereu de:



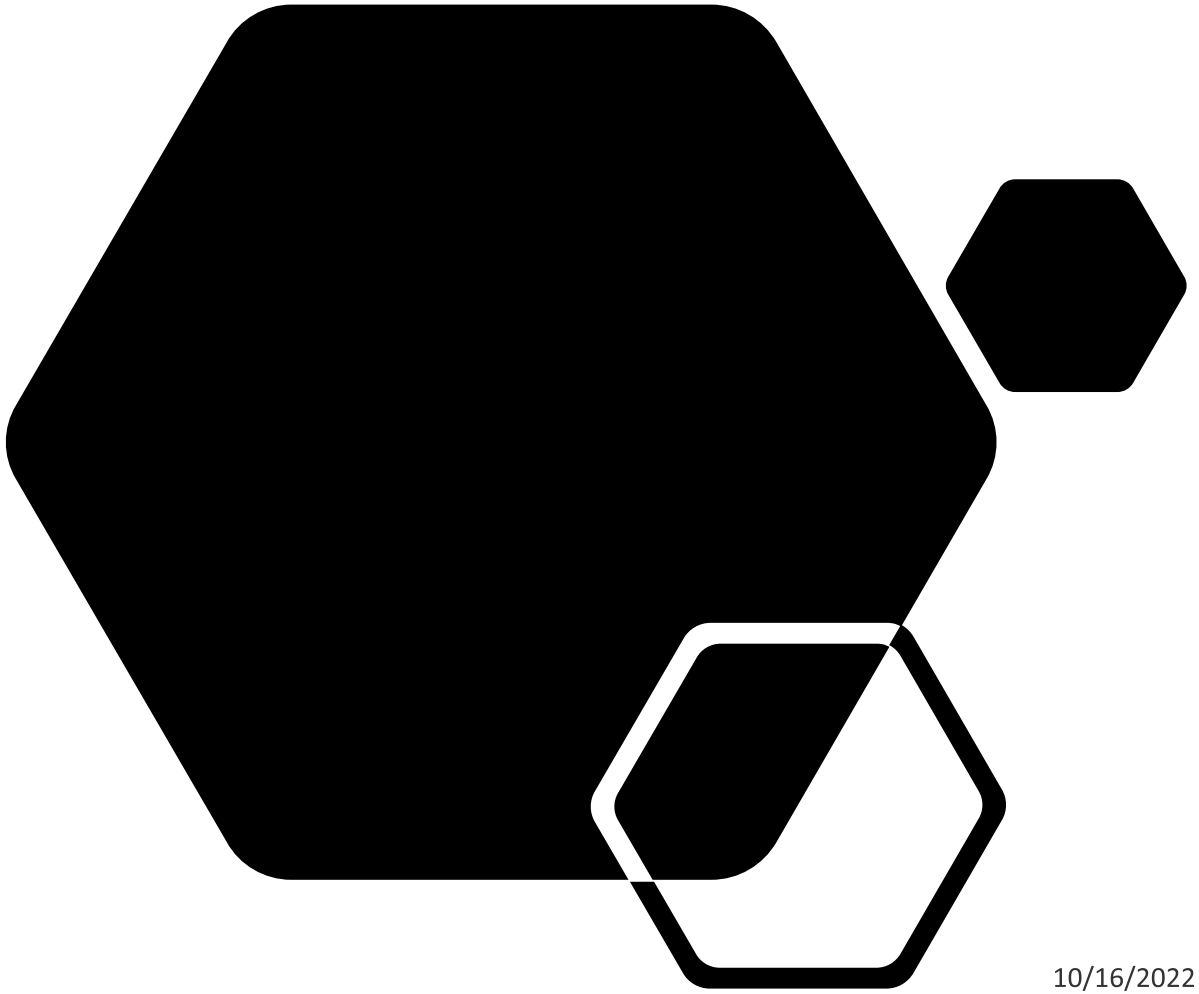
Distribuții

Indicatori

Distribuția normală

Tendință centrală

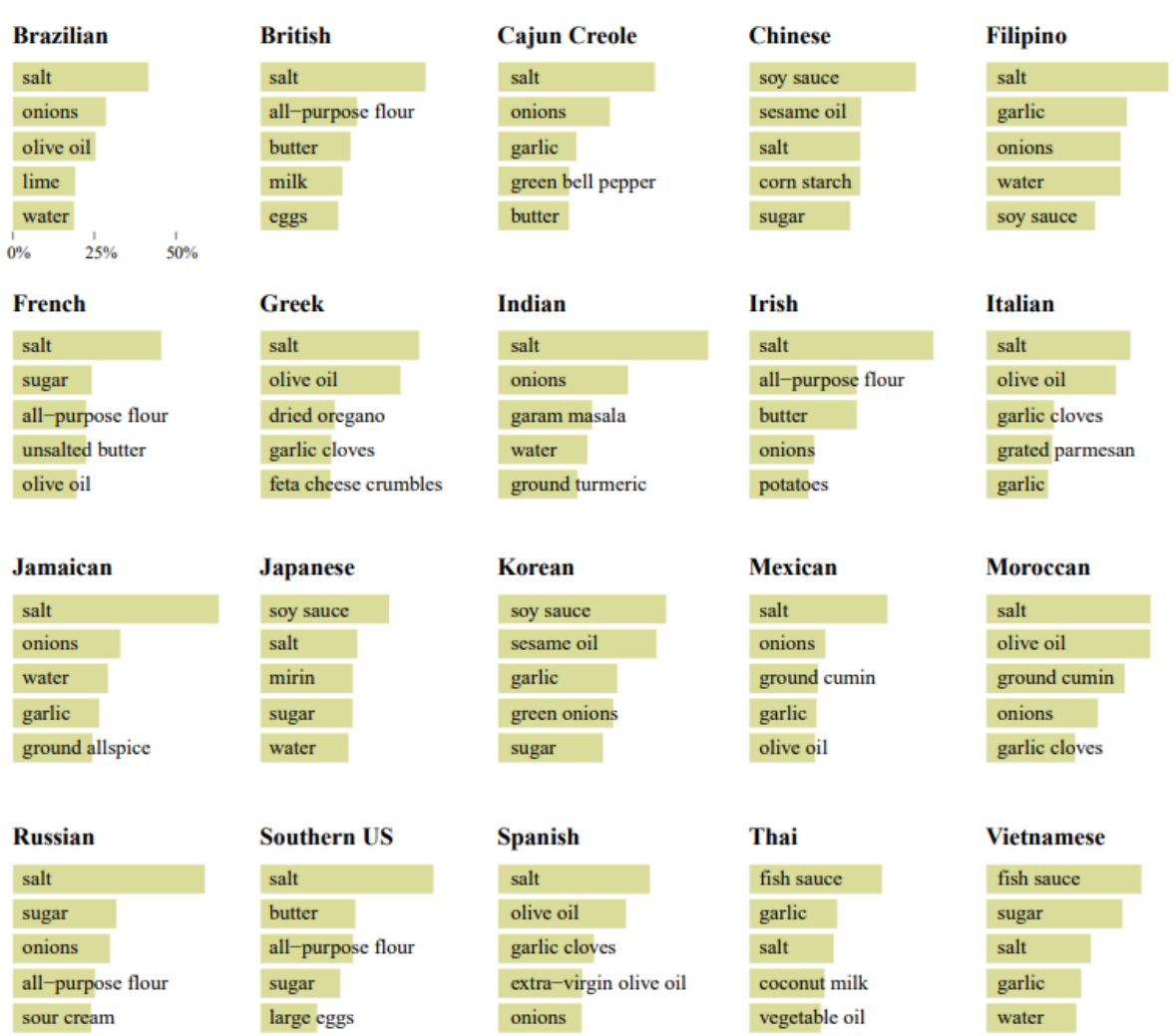
Variabilitate



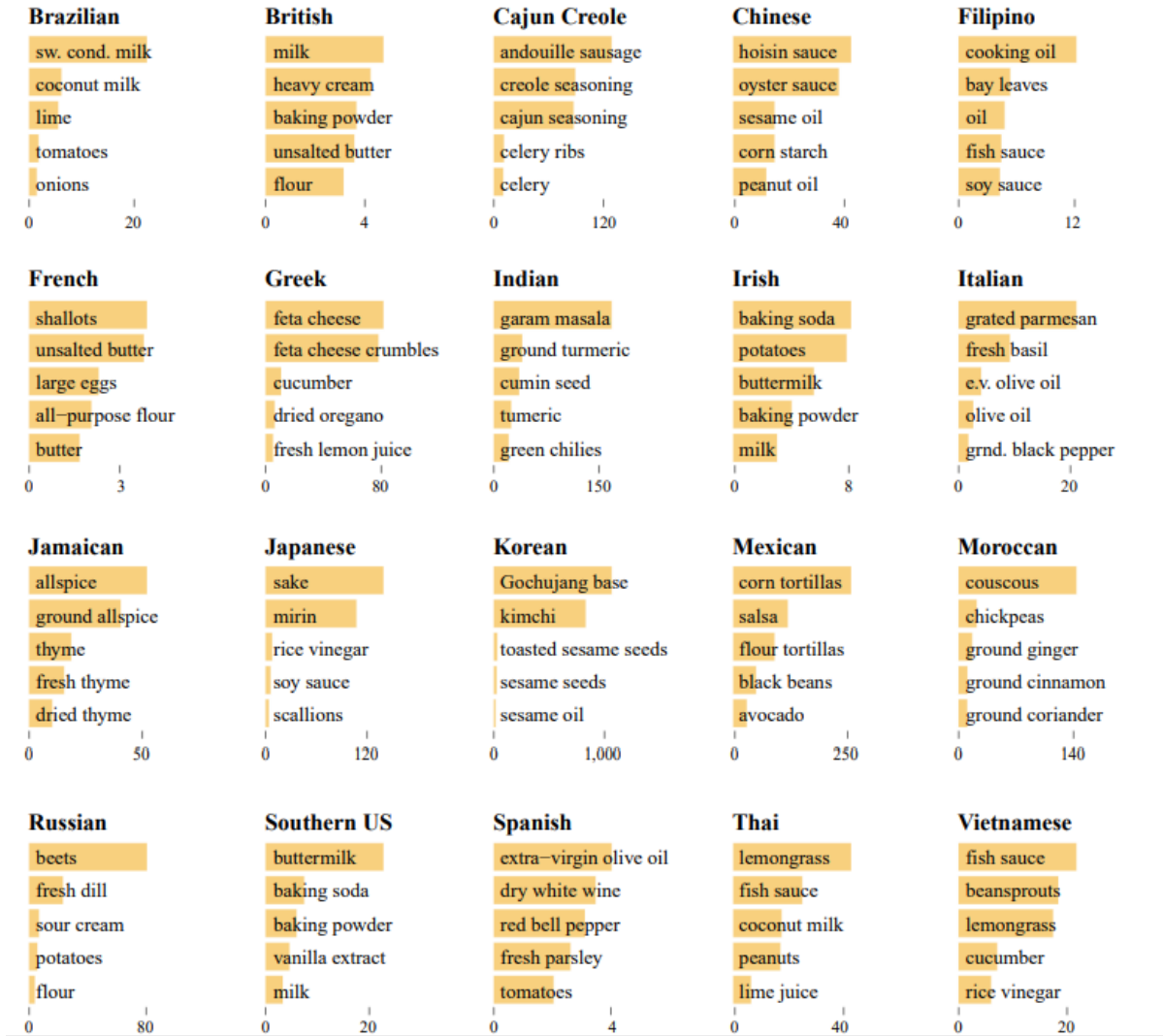
Indicatori: niveluri de măsurare

- 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)

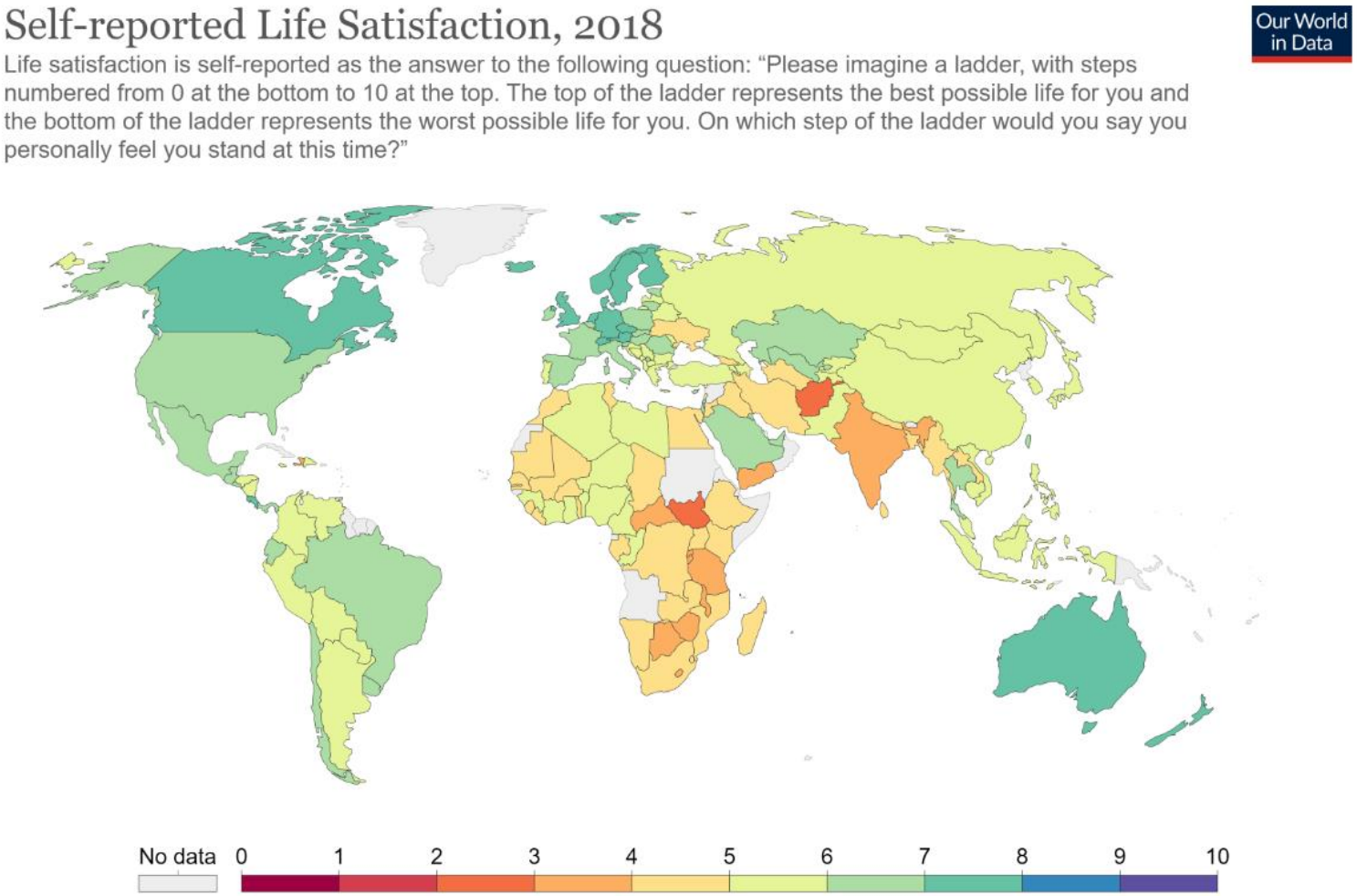
Most Used Ingredients



Most Cuisine-Specific Ingredients



Măsurare ordinală: satisfacția cu viața



Sursă

Distribuția normală

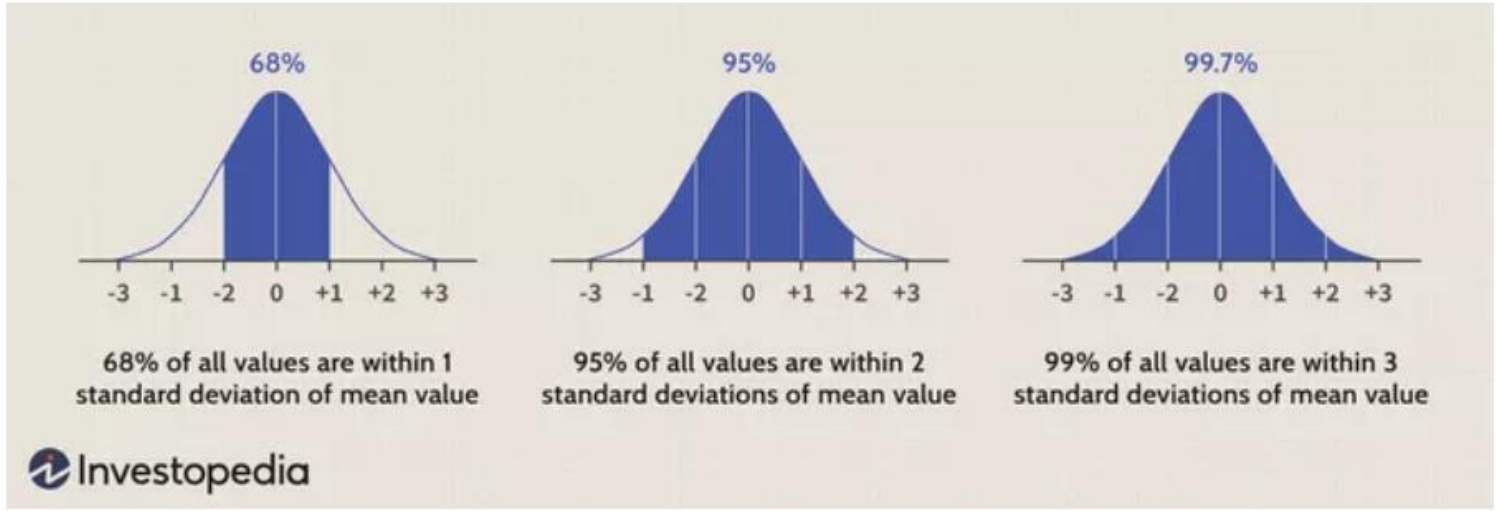
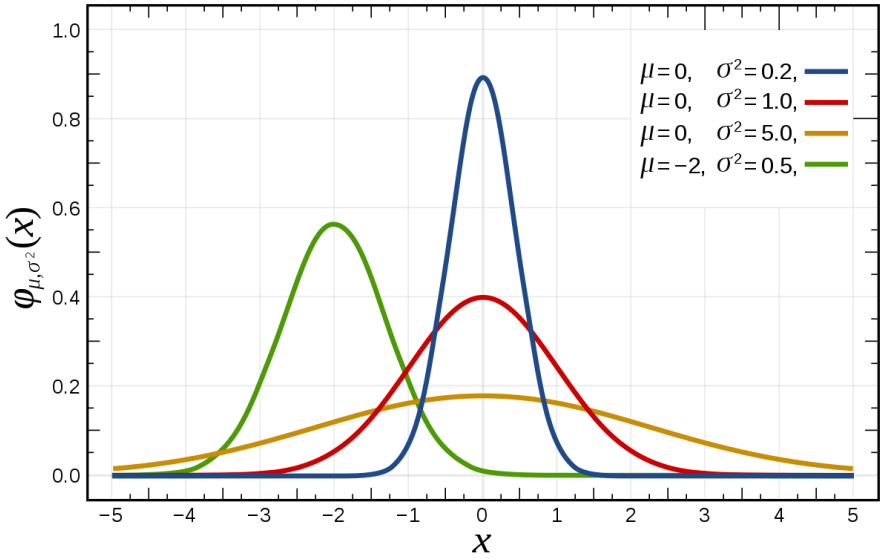


Image by Sabrina Jiang © Investopedia 2021

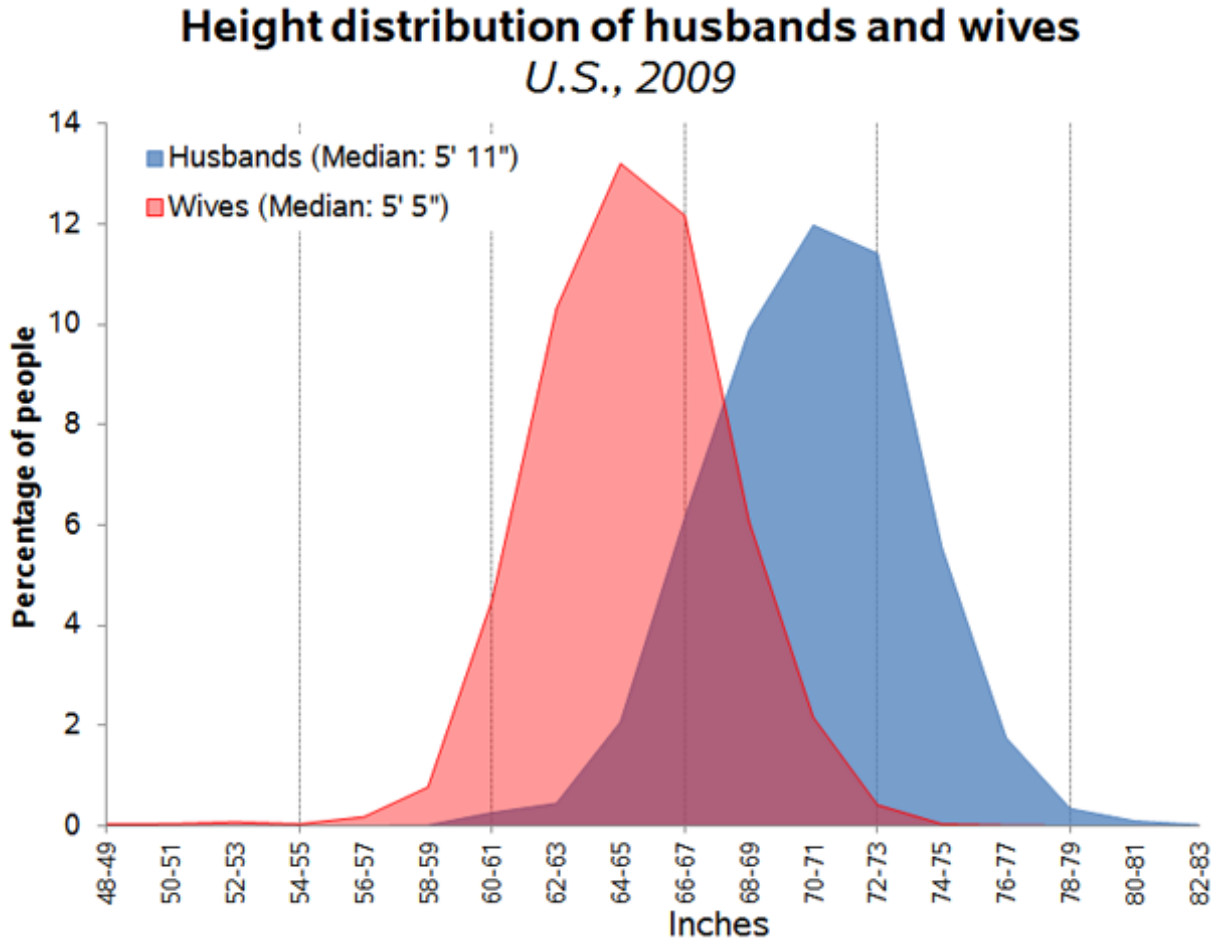


Stereotipuri vs. măsurare

- 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)
    - Media, mediana și modul
  - Variabilitate
    - Excepții, outliers
    - Percentile
    - Indici ai variabilității

Înălțimea și genul

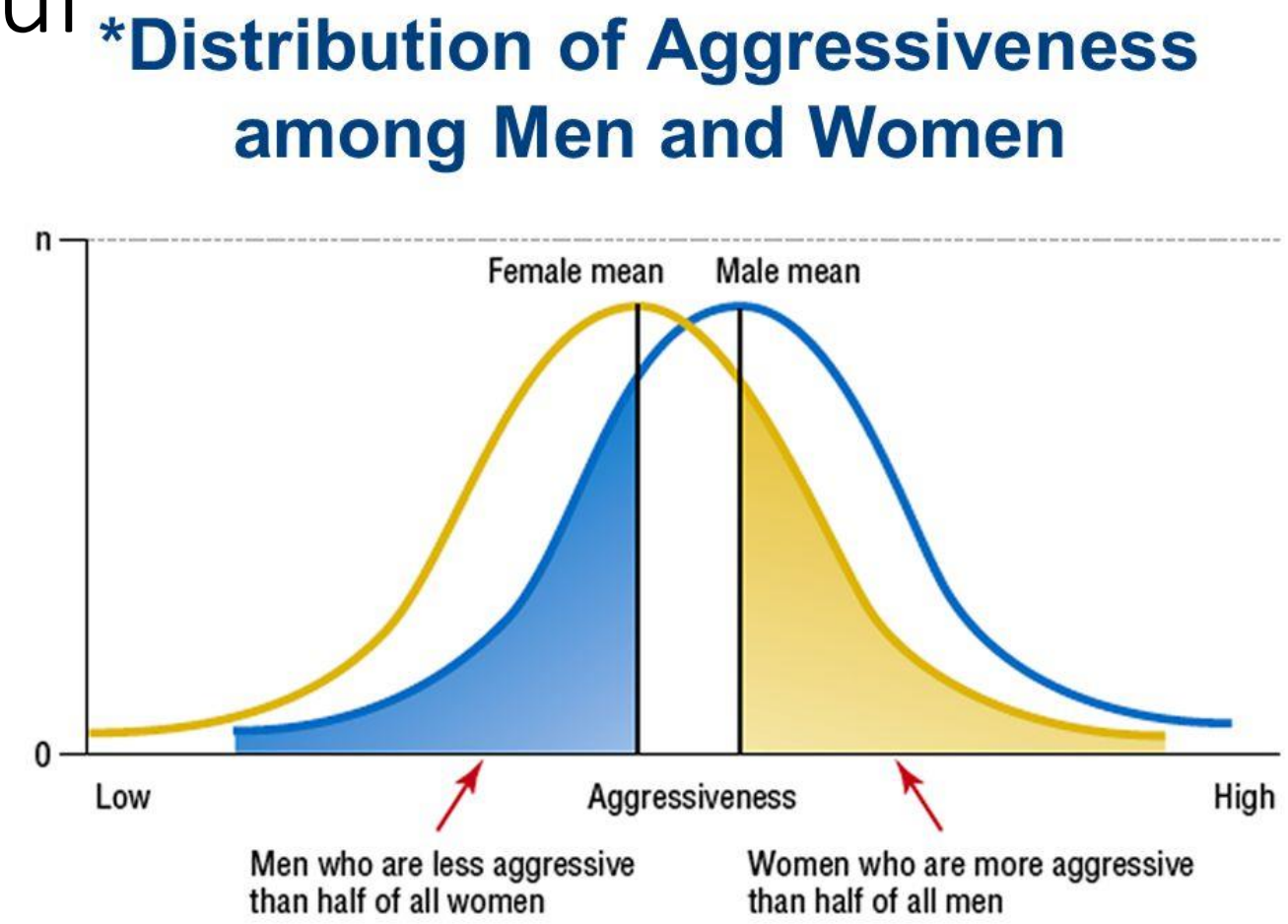
- Aproape întotdeauna soții sunt mai înalți decât soțiile
  - Cuplare selectivă
- Dar sunt întotdeauna bărbații mai înalți decât femeile?



The Atlantic

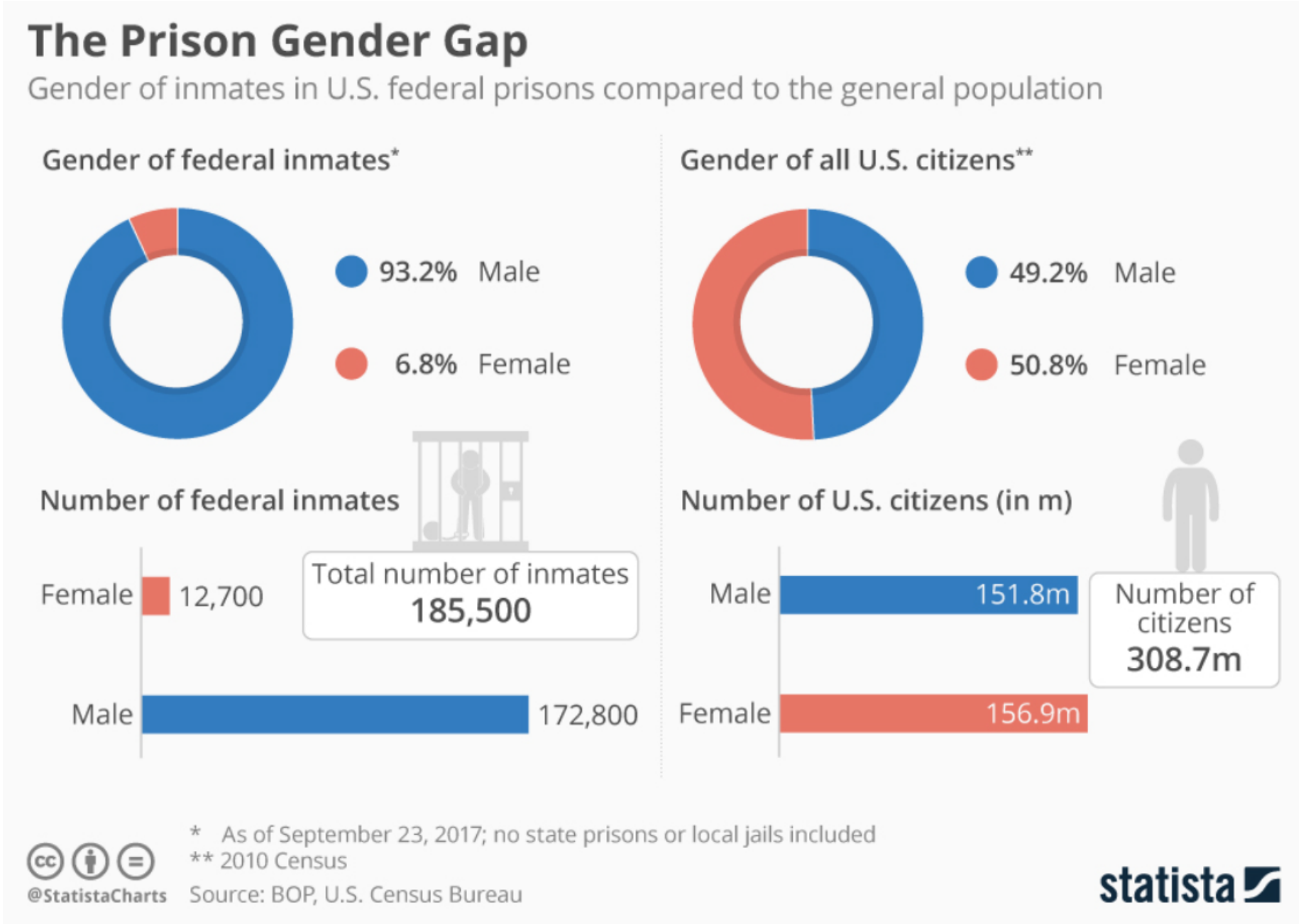
Agresivitatea și genul

- Închisorile dețin mult mai mulți bărbați decât femei pentru infracțiuni violente
- Rezultă că femeile nu sunt violente?
  - Distribuția violenței extreme
  - Discriminări ce influențează „măsurarea” violenței prin sentințe



Sursa

Sursă: Miller

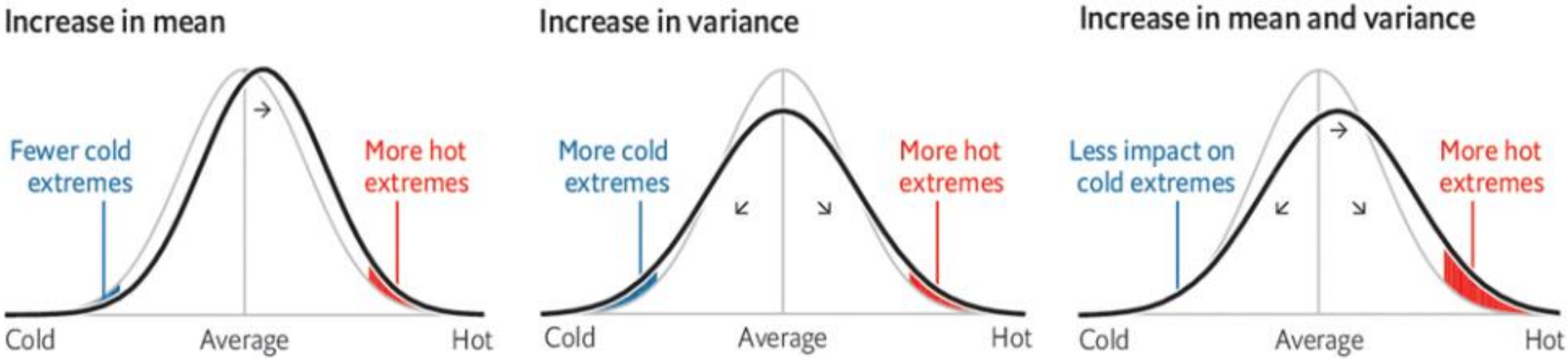

The Prison Gender Gap
Gender of inmates in U.S. federal prisons compared to the general population
Gender of federal inmates*
93.2% Male
6.8% Female
Gender of all U.S. citizens**
49.2% Male
50.8% Female
Number of federal inmates
Female 12,700
Male 172,800
Total number of inmates
185,500
Number of U.S. citizens (in m)
Male 151.8m
Female 156.9m
Number of citizens
308.7m
* As of September 23, 2017; no state prisons or local jails included
** 2010 Census
@StatistaCharts
Source: BOP, U.S. Census Bureau
statista

Sursa

Schimbarea climatică și extremele

→ A hotter planet is a more extreme one; some regions become unliveable

Effect of changes in global temperature



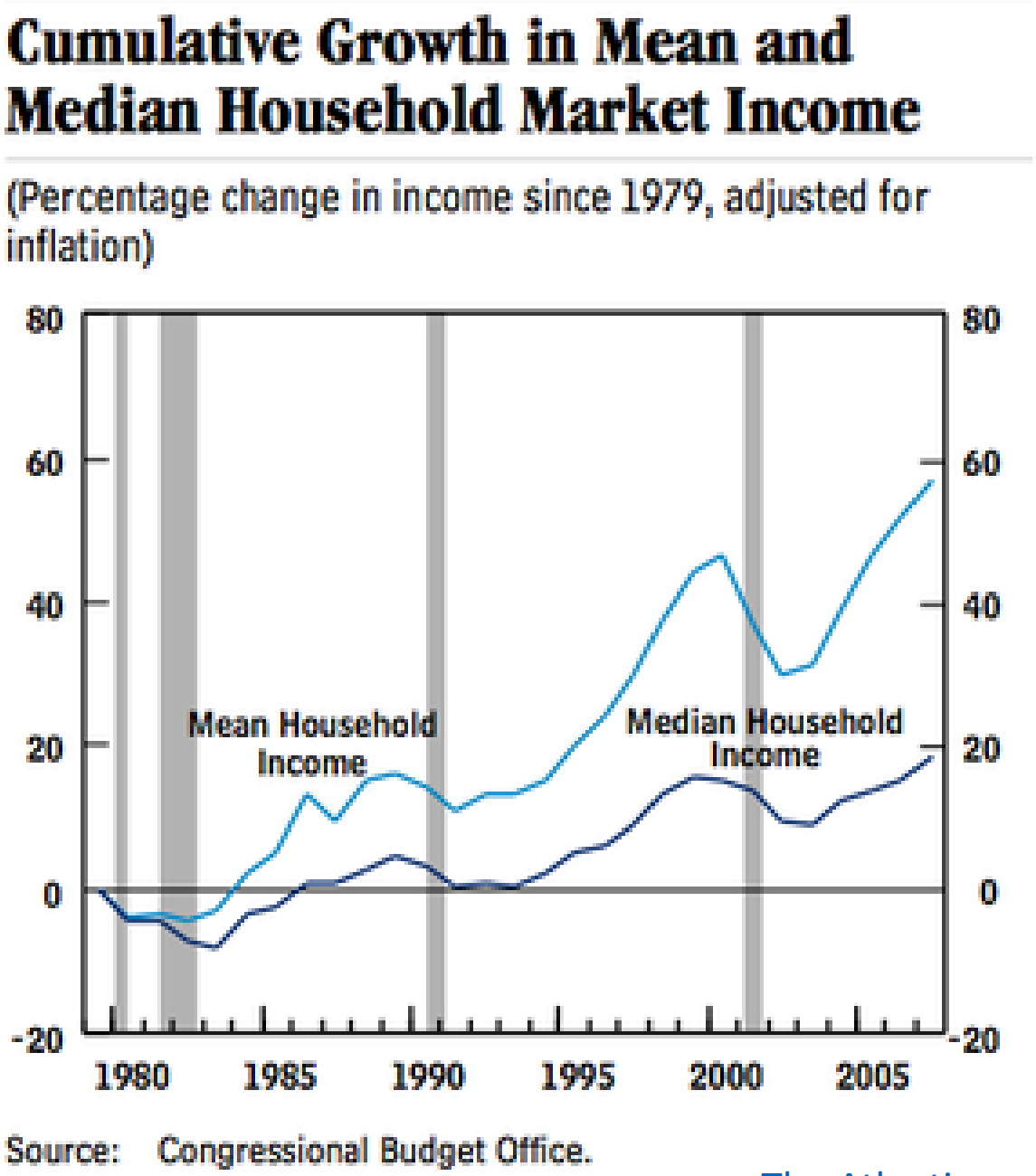
Source:
The Economist

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 $\sigma_X^2 = \frac{\sum_{i=1}^{n}(x_i - \mu_X)^2}{n}$
  - Abaterea standard $\sigma_X = \sqrt[2]{\sigma_X^2}$

Tendința centrală: medie vs. mediană

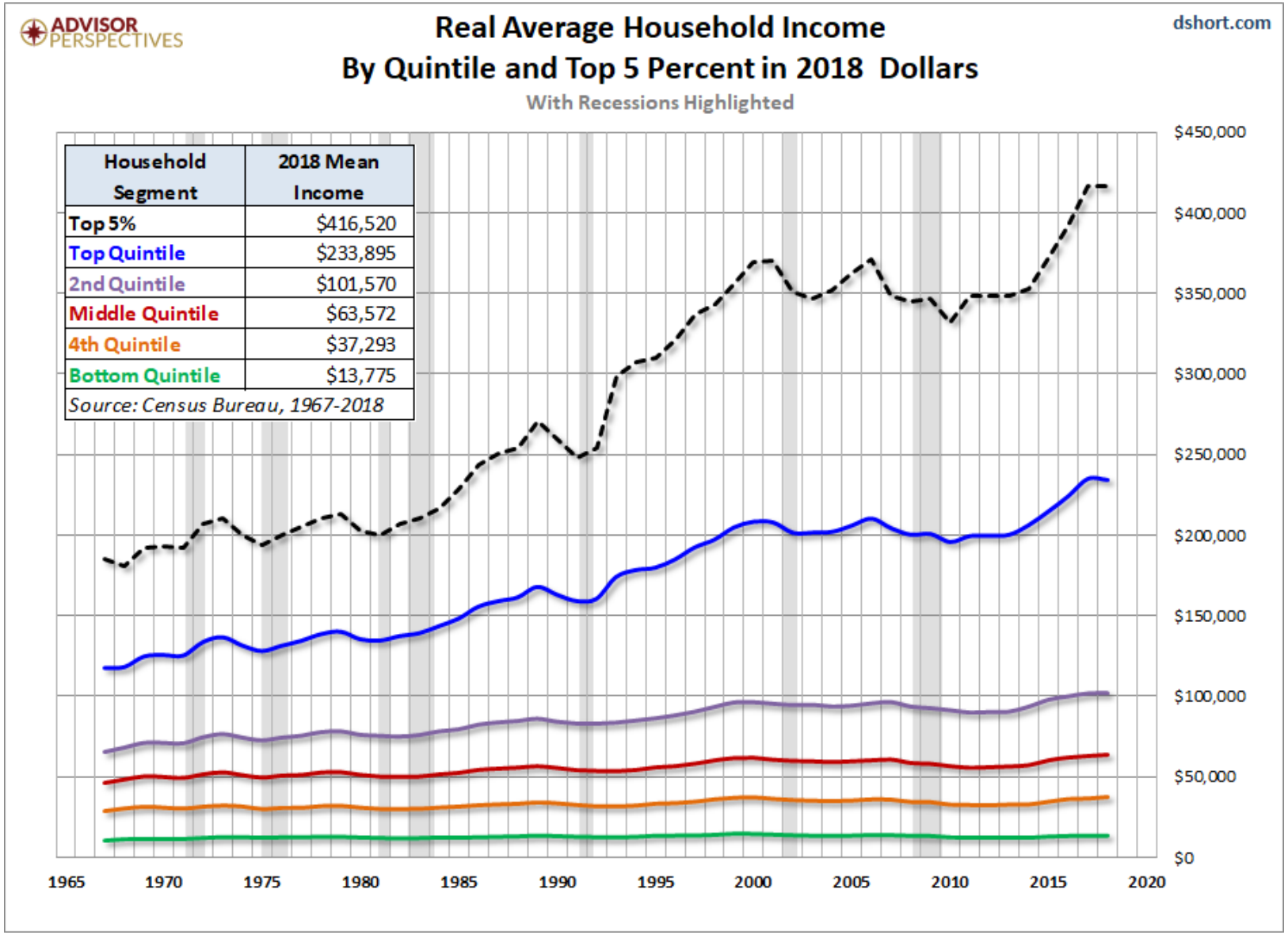
- Media este mai influențată de valorile extreme
- Extremele tip top 1% trag în sus media averii gospodăriilor



The Atlantic

Percentile:

Înțelegerea dinamicilor reale a creșterii venitului

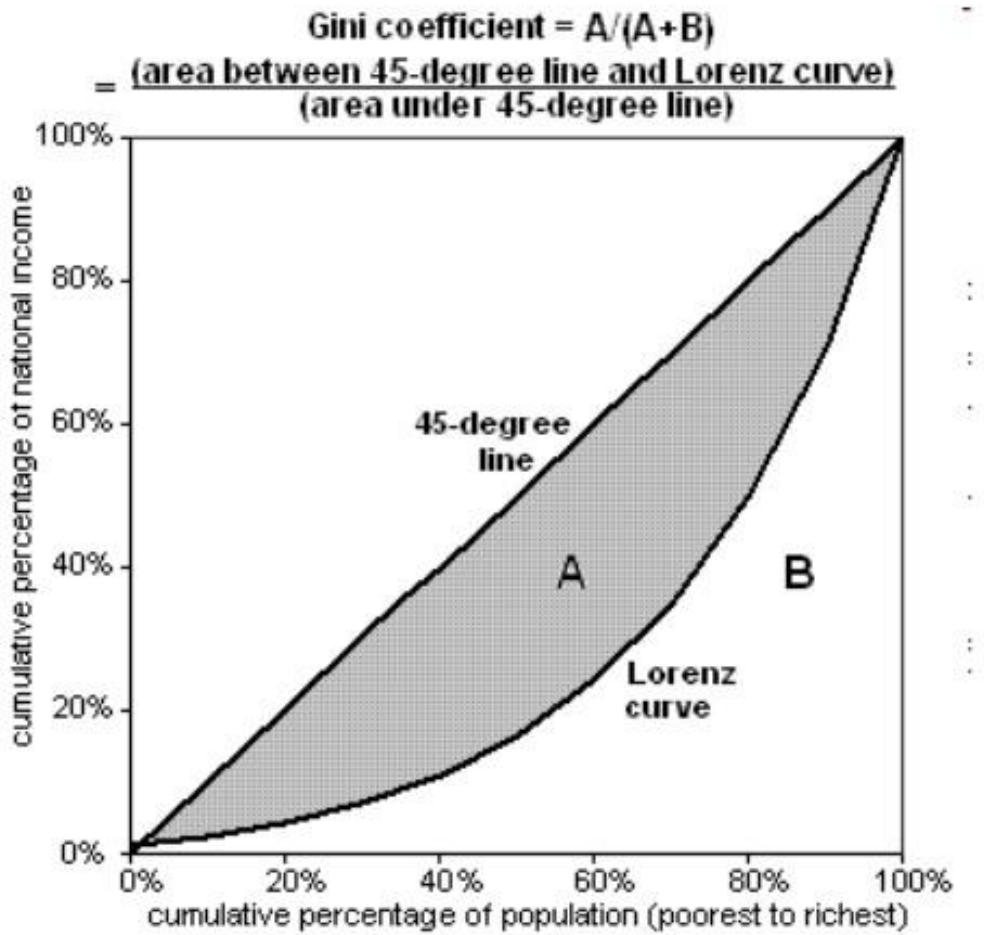


Sursă

Variabilitatea averii:

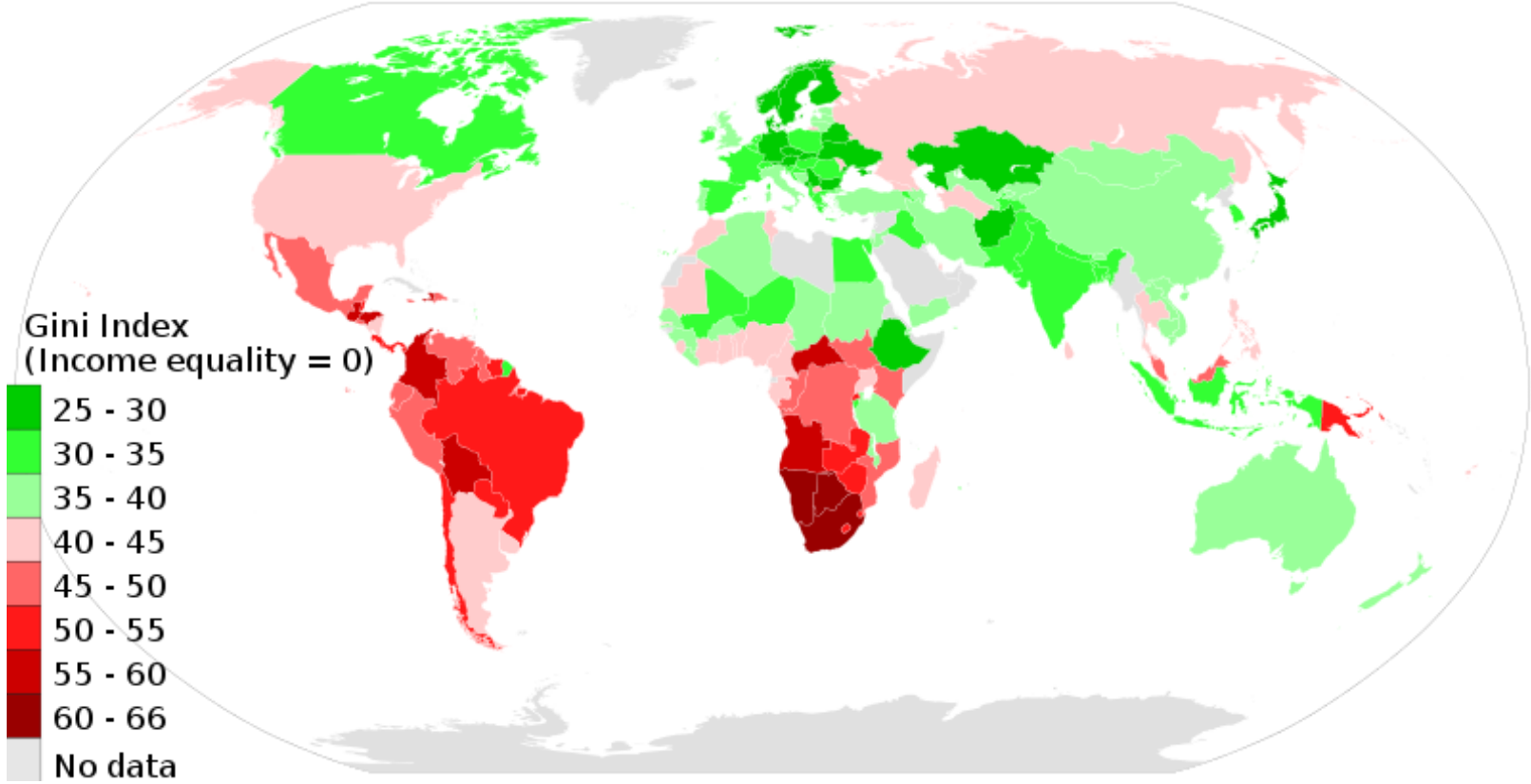
Coeficientul Gini

Gini Coefficient



2

Variabilitatea variabilității averii

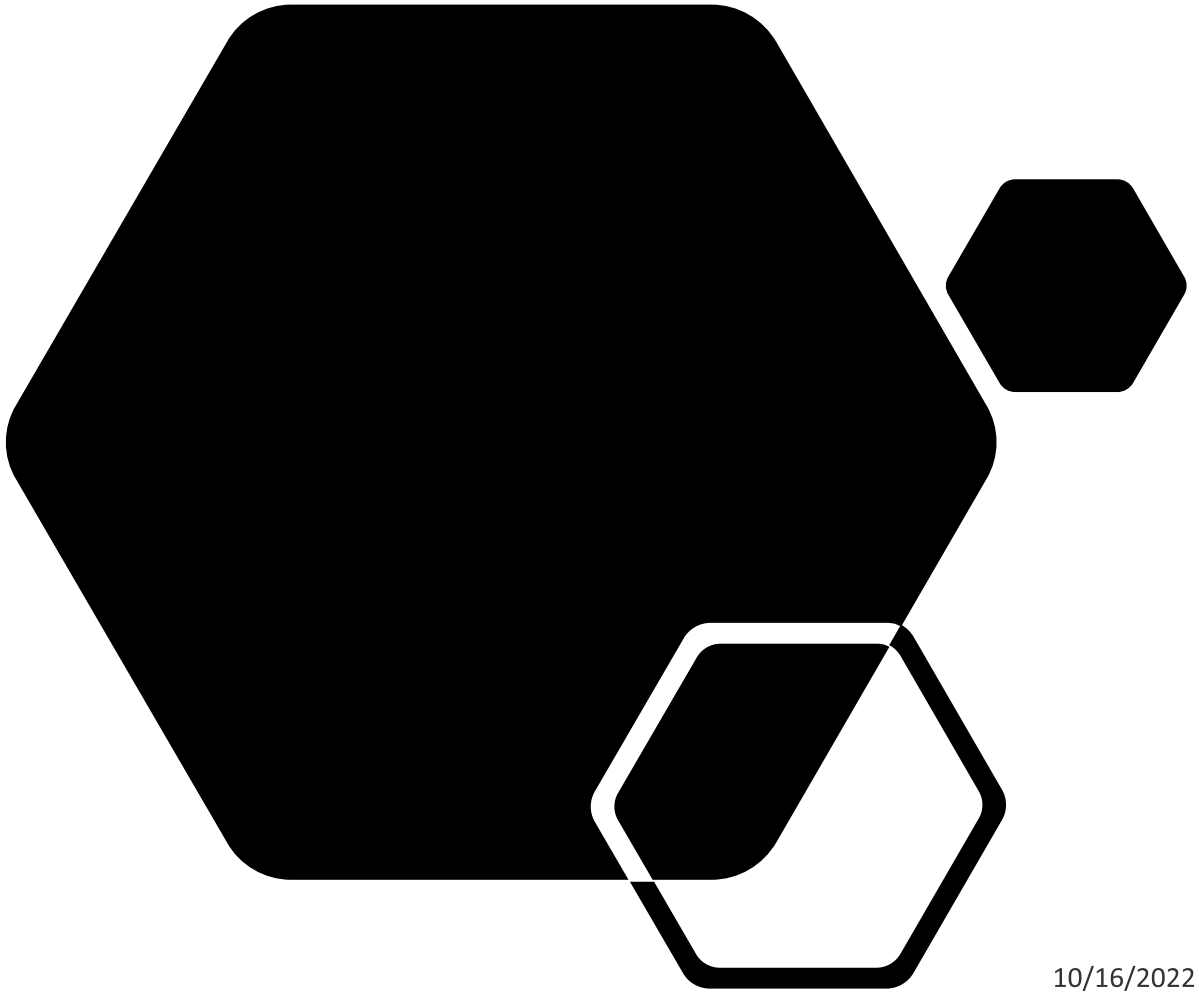


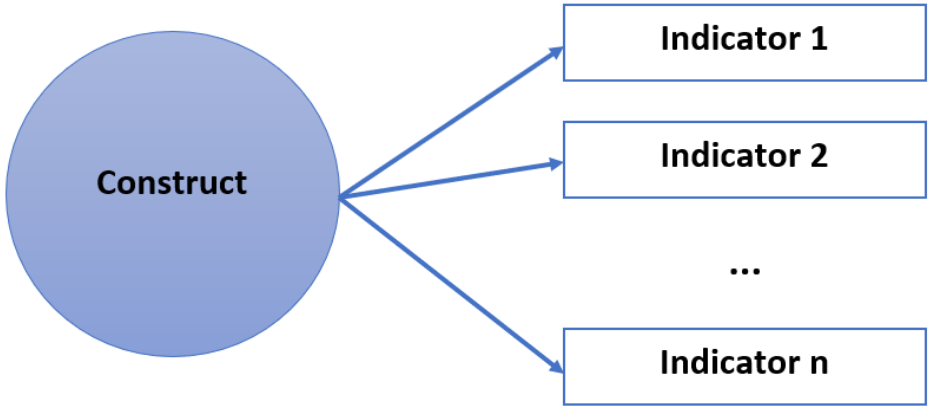
Wikimedia

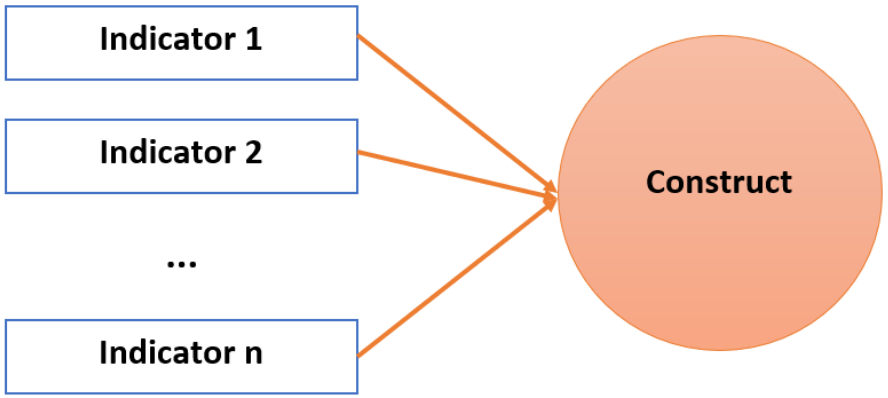
Modele de măsurare

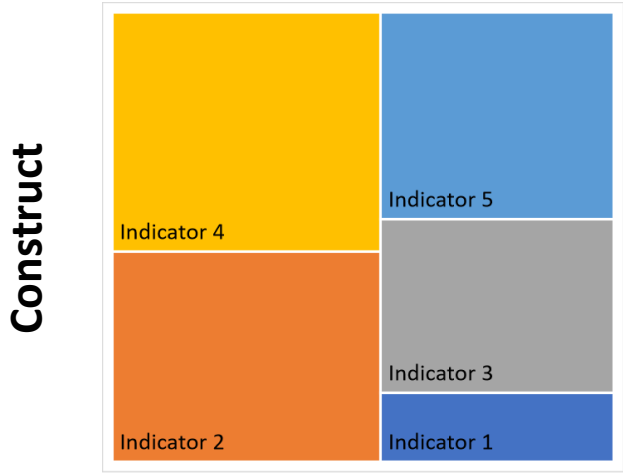
Modelul reflectiv

Modelul formativ









Modelul reflectiv:
Constructul este cauza indicatorilor
Indicatorii corelează puternic
Analiză factorială: identificarea factorului latent

Modelul formativ:
Indicatorii cauzează / compun constructul
Indicatorii pot să nu coreleze
Sumă / medie ponderată etc

Exemple

- Ce măsoară notele? Trebuie notele să fie normalizate?

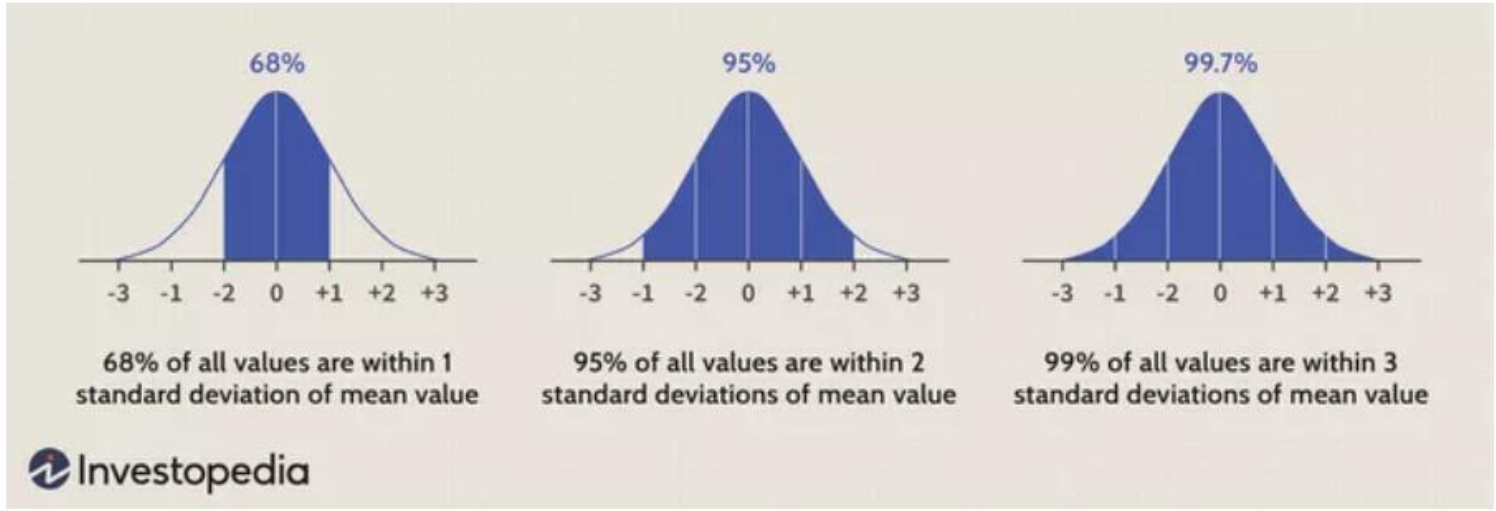
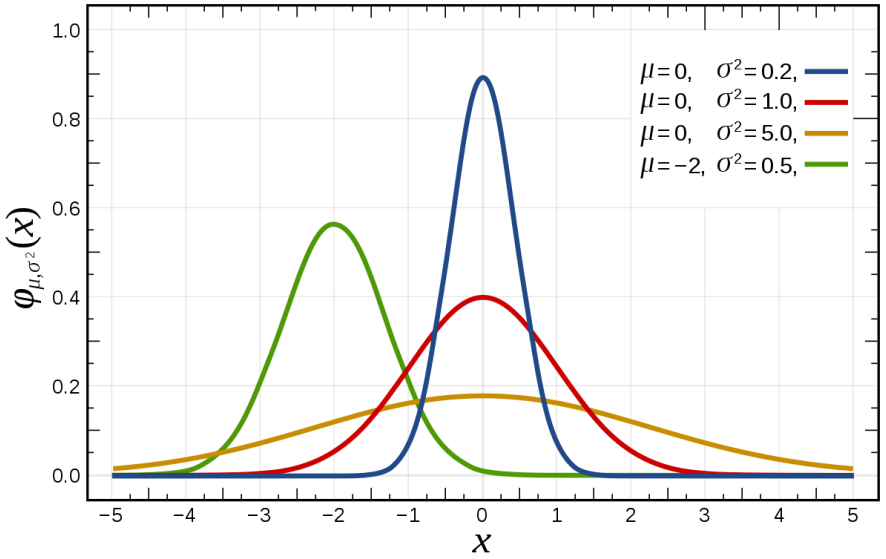
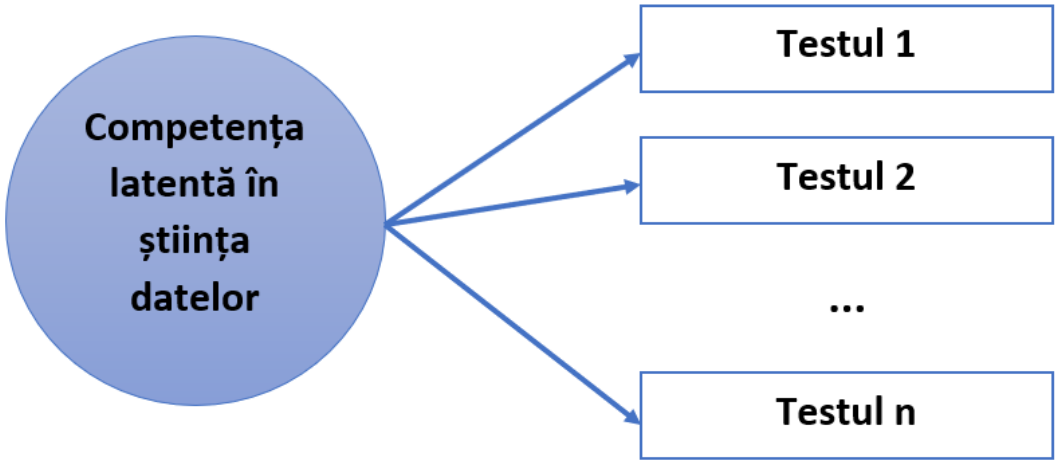
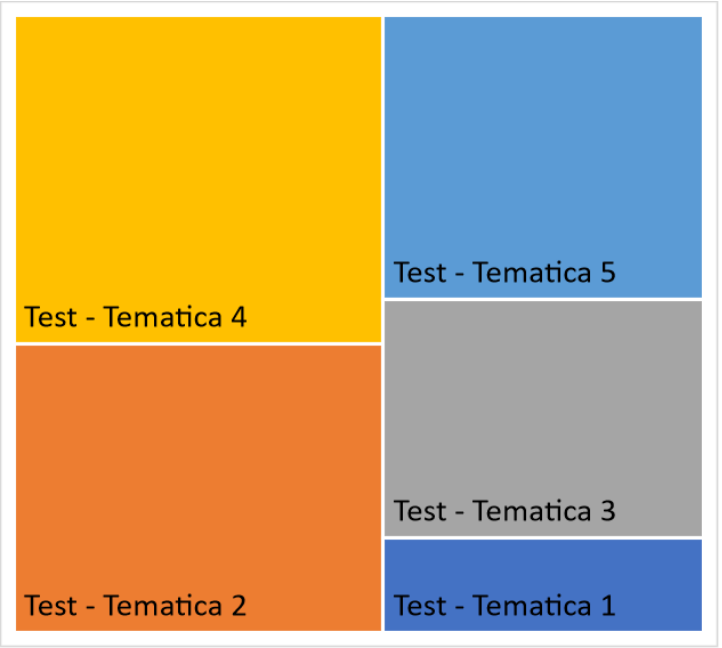


Image by Sabrina Jiang © Investopedia 2021







Model reflectiv – competența latentă

- Constructul este cauza indicatorilor
- Testele reflectă competența latentă (invizibilă) în știința datelor
- Toate testele măsoară imperfect aceeași competență
- Notele la teste sunt corelate (imperfect)
- Notele sunt distribuite normal

Model formativ – cunoștințe

- Indicatorii compun constructul
- Testele măsoară fiecare o altă tematică
- Cele 5 tematici acoperă cunoștințele de știința datelor
- Notele la teste pot să nu fie corelate
- Notele pot fi distribuite variabil (eg toți studenții au cunoștințele cerute)

Ce modele de măsurare au...

- Inflația
- Inteligența (IQ)

Ce este inflația?

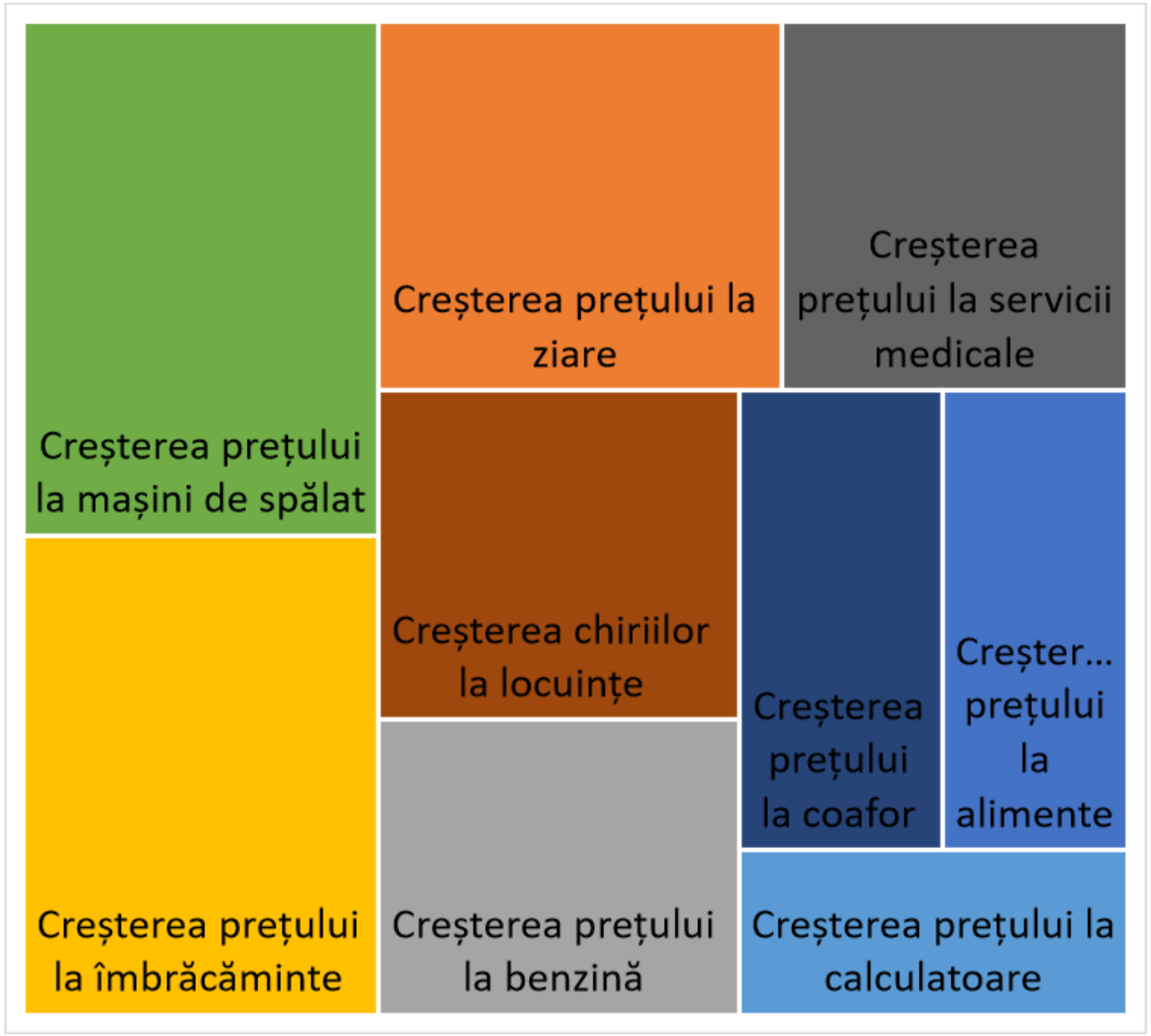
„Coșul” de bunuri de consum

Fiecare produs din acest coș are un preț care poate varia în timp.

Rata anuală a inflației este dată de prețul coșului integral într-o anumită lună comparat cu prețul acestuia în aceeași lună a anului precedent.

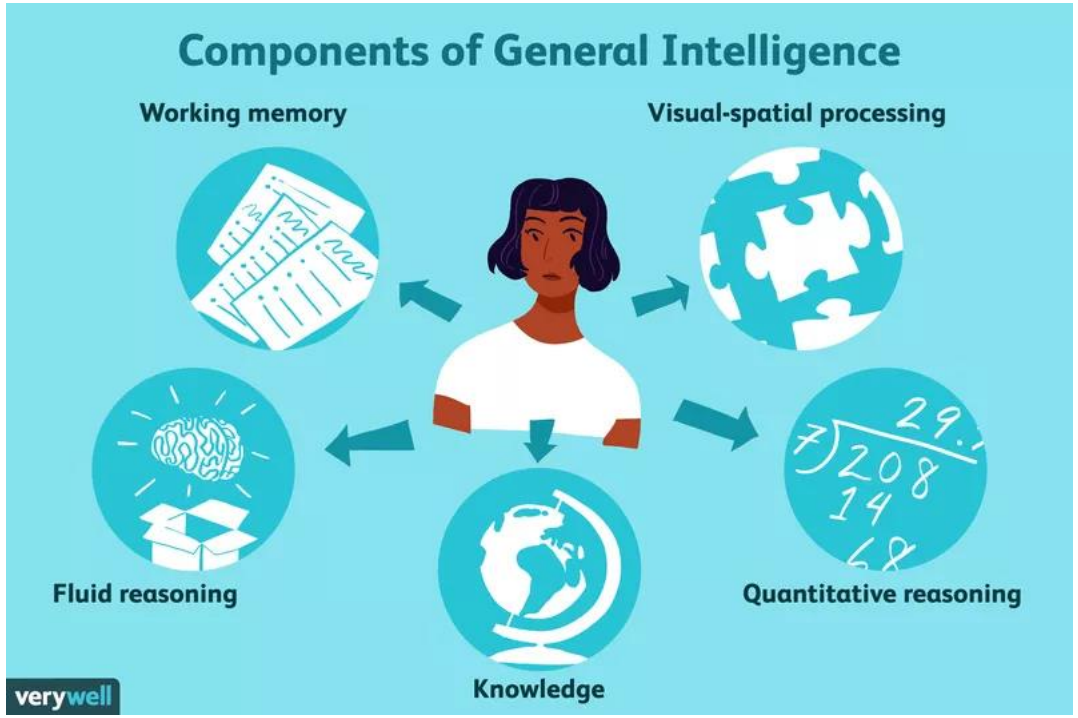
Ponderi diferite pentru categorii diferite

- Ponderi mari pentru produsele pentru care cheltuim mai mult

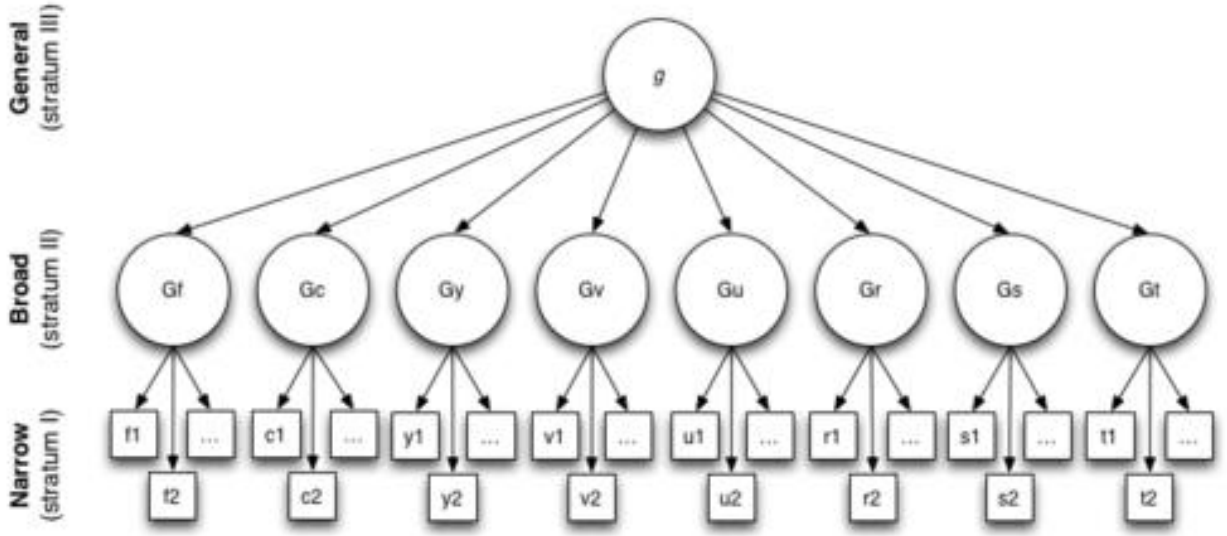


Sursă

Ce este IQ?



Sursă



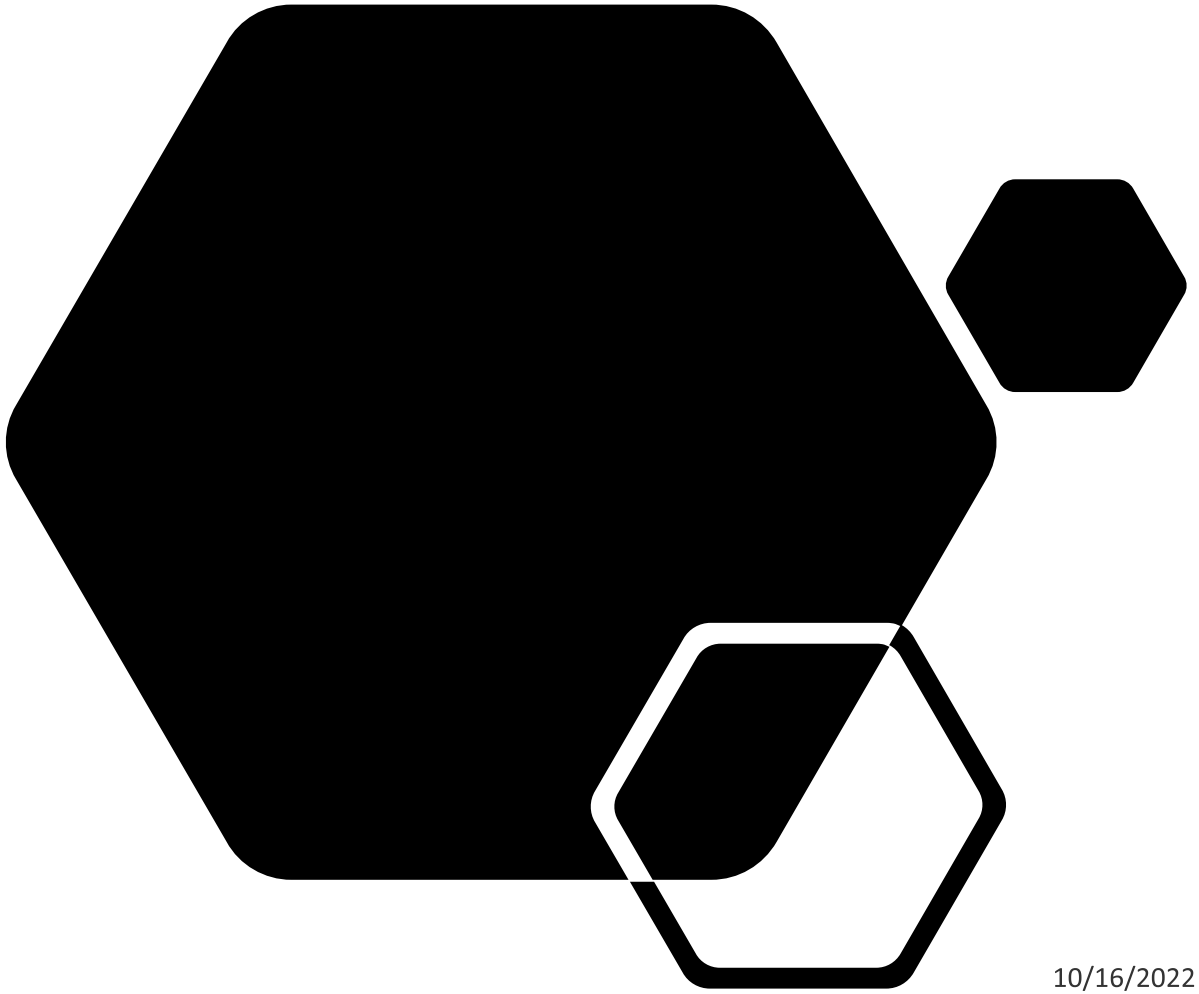
An illustration of John B. Carroll's three stratum theory, an influential contemporary model of cognitive abilities. The broad abilities recognized by the model are fluid intelligence (Gf), crystallized intelligence (Gc), general memory and learning (Gy), broad visual perception (Gv), broad auditory perception (Gu), broad retrieval ability (Gr), broad cognitive speediness (Gs), and processing speed (Gt). Carroll regarded the broad abilities as different "flavors" of *g*.

Sursa

Erori de măsurare

Erori sistematice și aleatorii

Validitate și fidelitate



Precizia măsurării

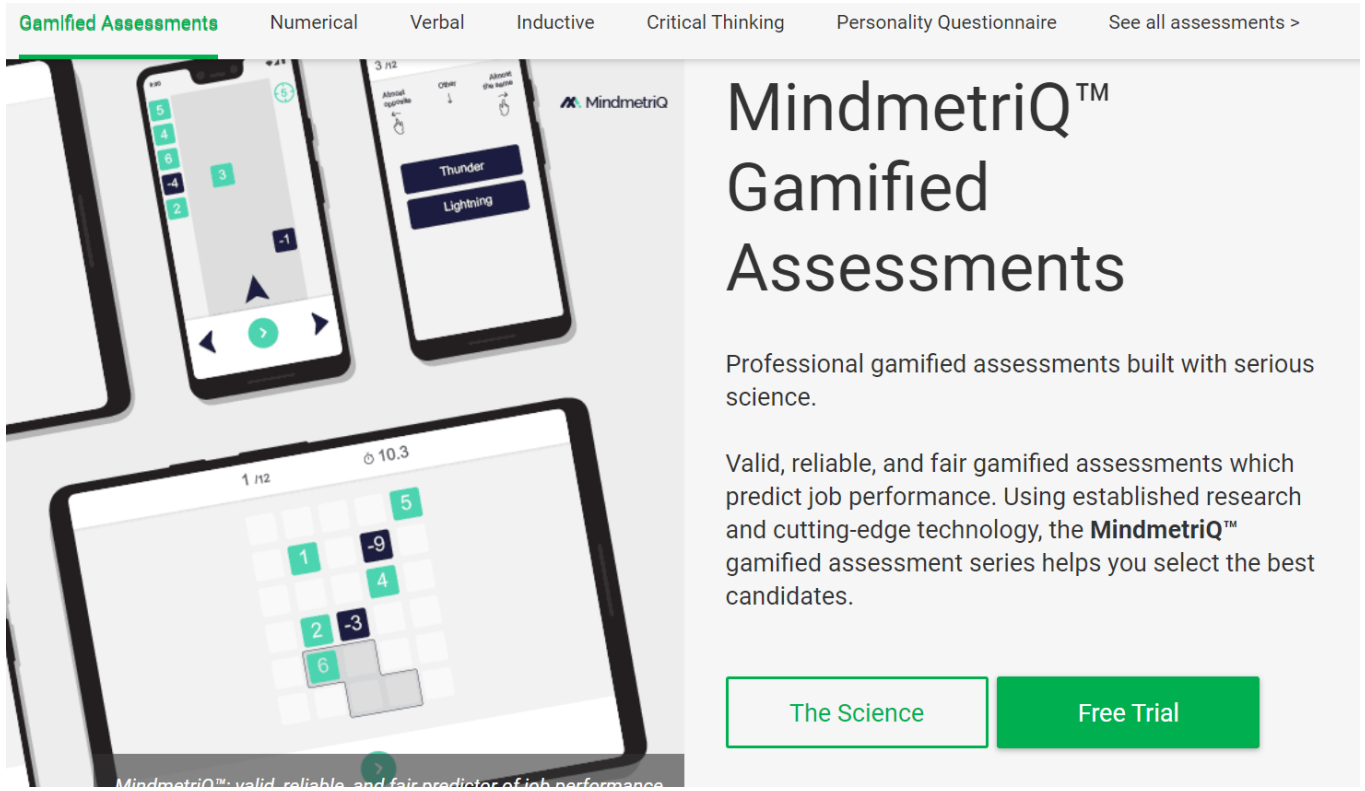
- En. *accuracy*
- Valoarea măsurată - valoarea reală =eroare
- Tipuri de erori în funcție de direcția lor:
  - Erori **aleatoare**
  - Erori **sistematice**: dezirabilitatea socială
    - Orientarea sexuală
    - Preferințele politice
- Măsurarea timpului:
  - Ochiometric, clepsidră: erori aleatoare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

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: 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** prin **data nașterii**

- Foarte precis
- Foarte repetabil
- Cât de valid?
  - Este data nașterii un indicator valid pentru constructul „zodie”?

Scenarii

accurate
repeatable

Temperatura [termometrul]
- Erori aleatorii reduse

not accurate
repeatable

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

accurate
not repeatable

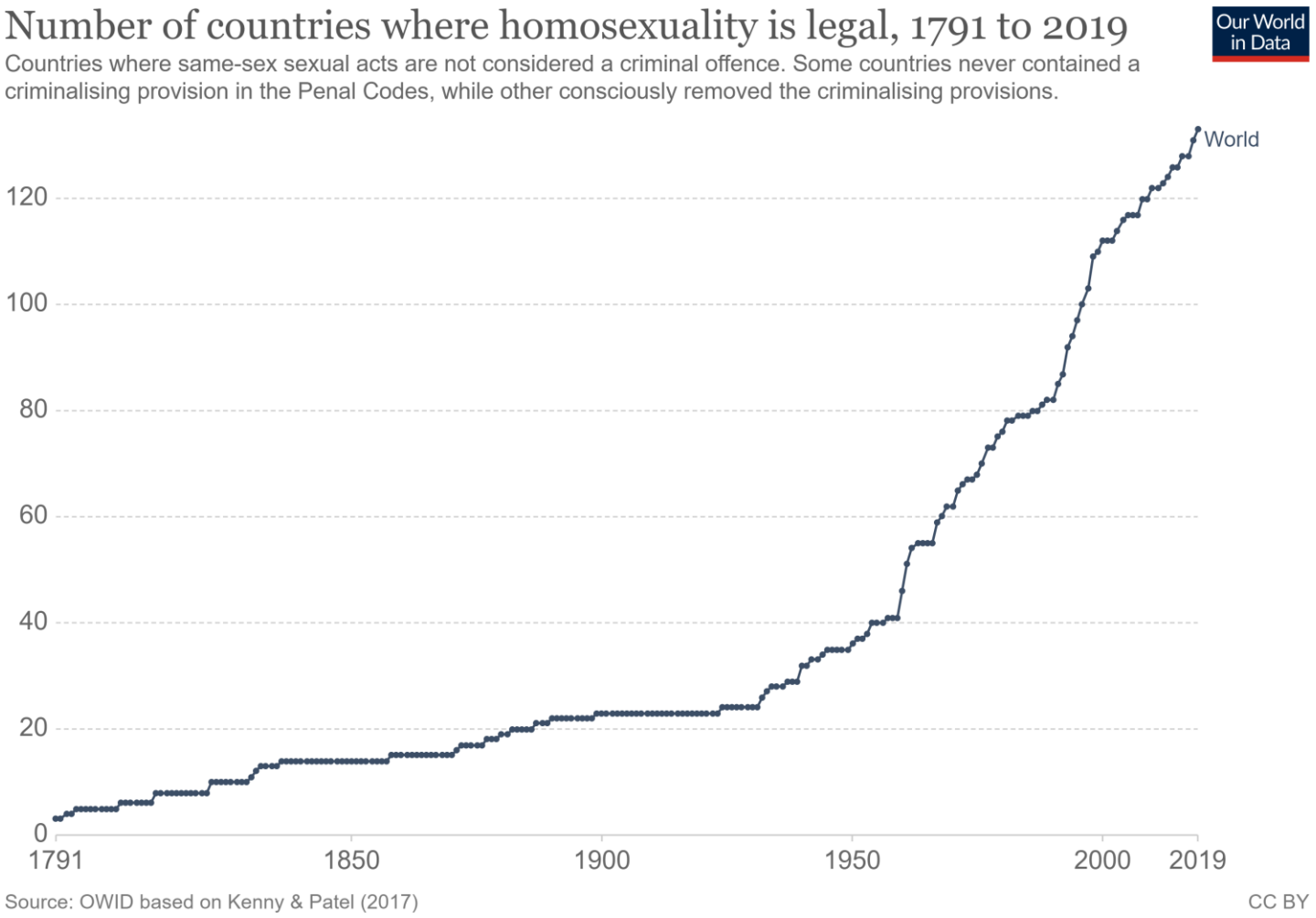
Ora
[capul pe geam]
- Erori aleatorii mari

not accurate
not repeatable

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?

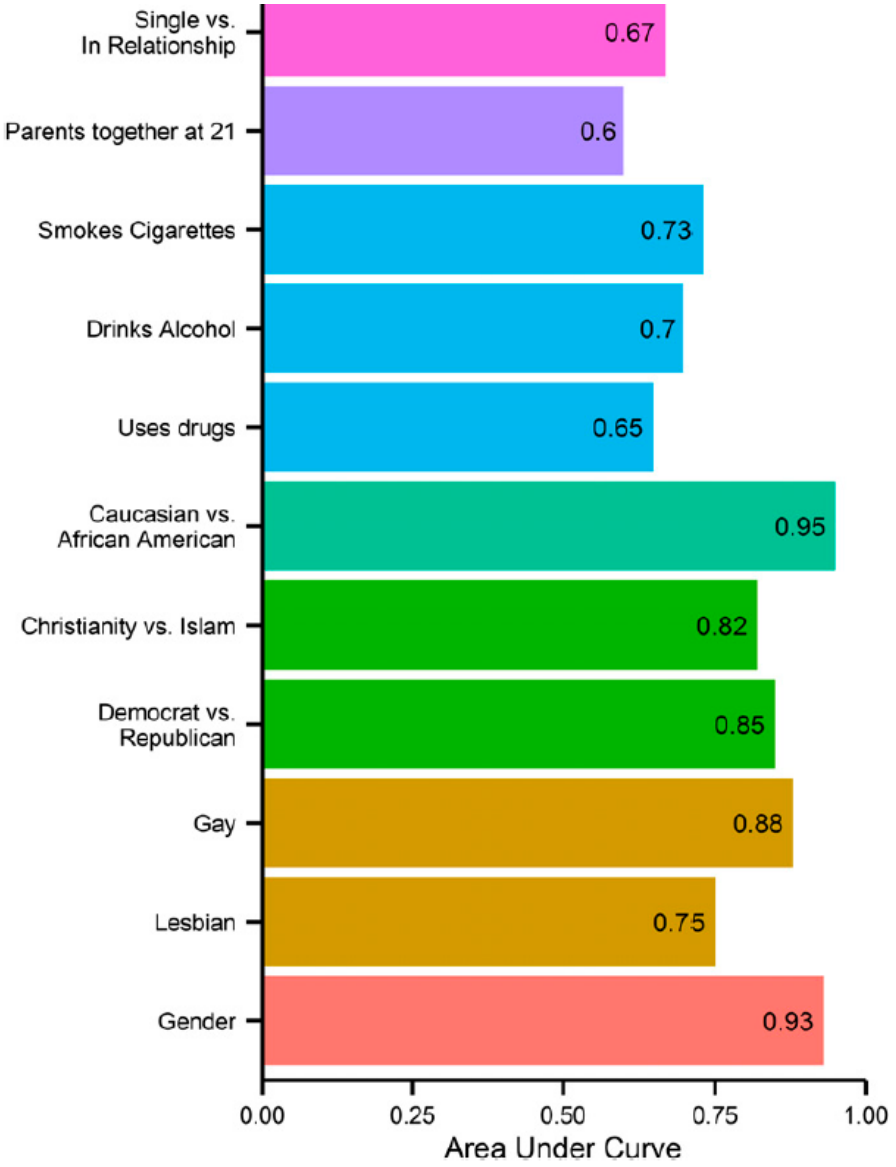
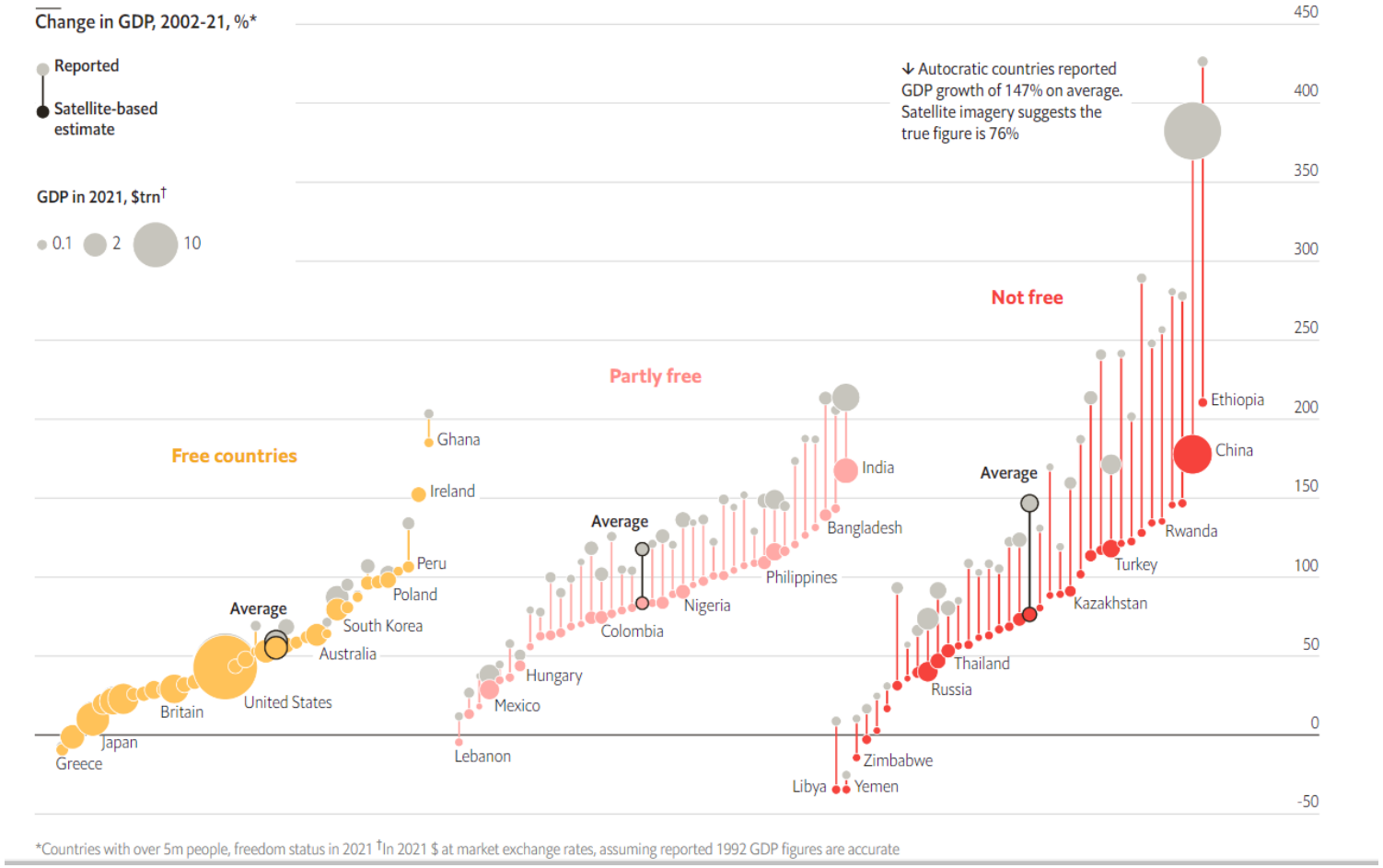


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

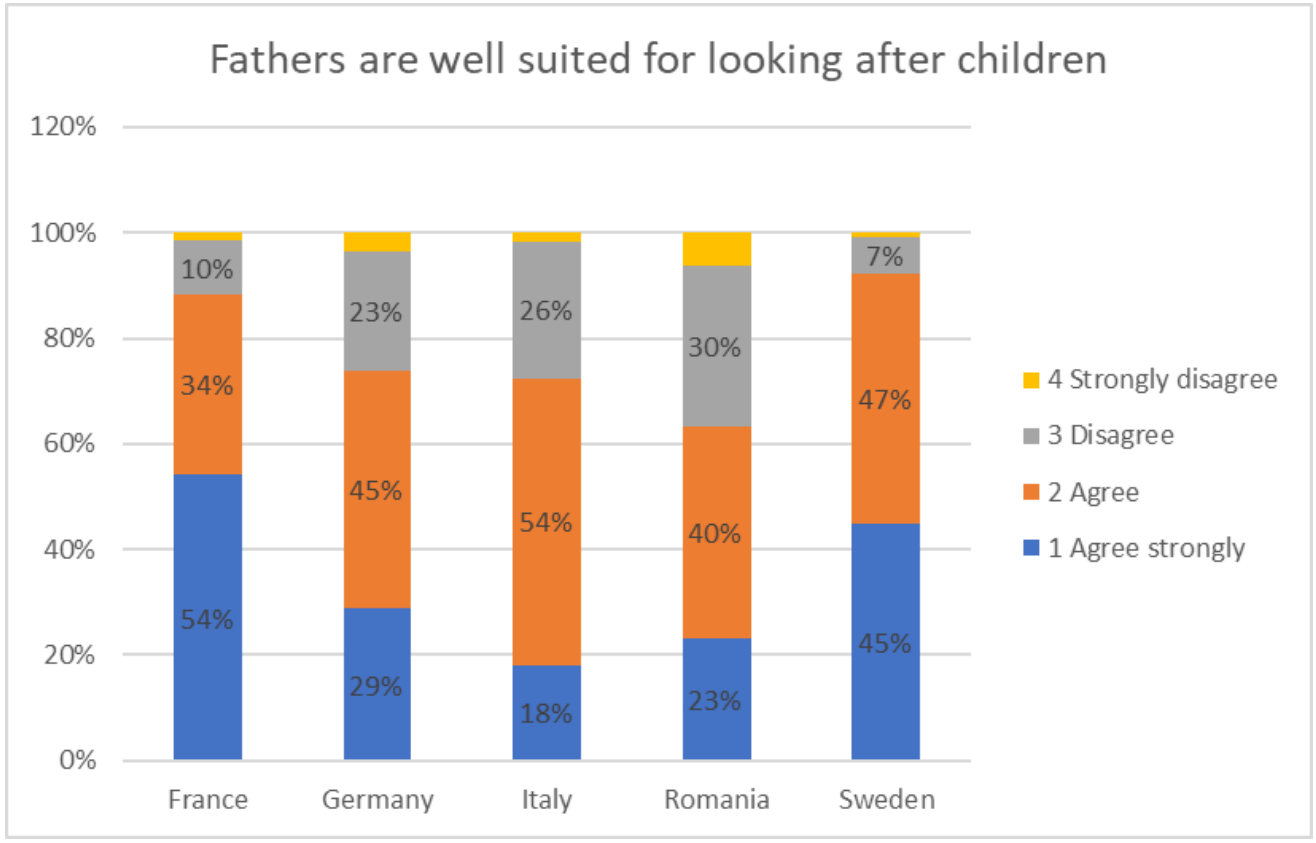
Măsurarea creșterii GDP în state democratice sau autoritare



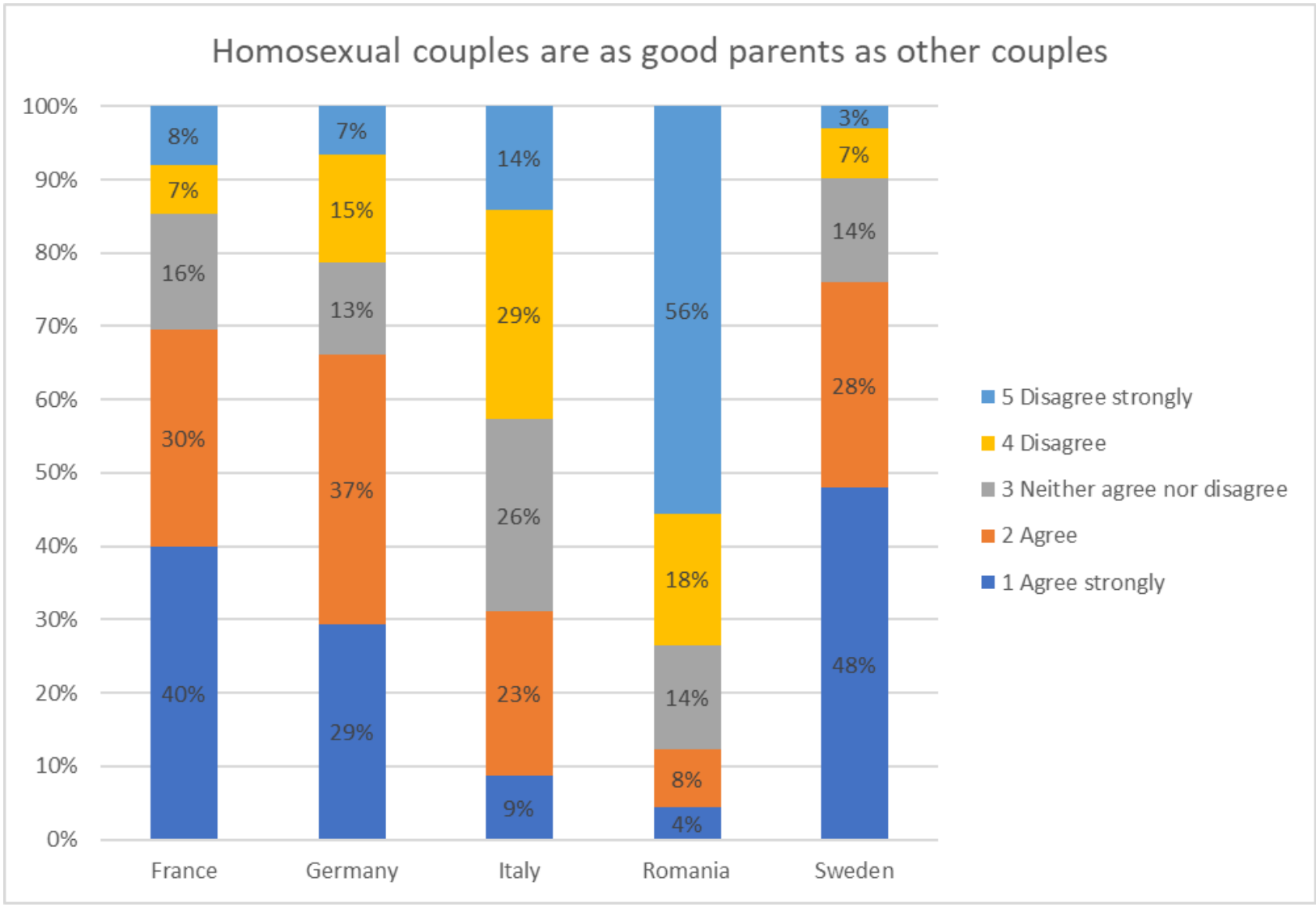
Sursa:
The Economist

Alte tipuri de erori pentru măsurarea opiniei

- Stilul de răspuns
  - Preferința pentru a răspunde „Da” (agreeabilitate, aquiescență)
  - Preferința pentru a răspunde „Nu știu”
- Alegerea scalei
  - Cu punct de mijloc sau fără punct de mijloc?



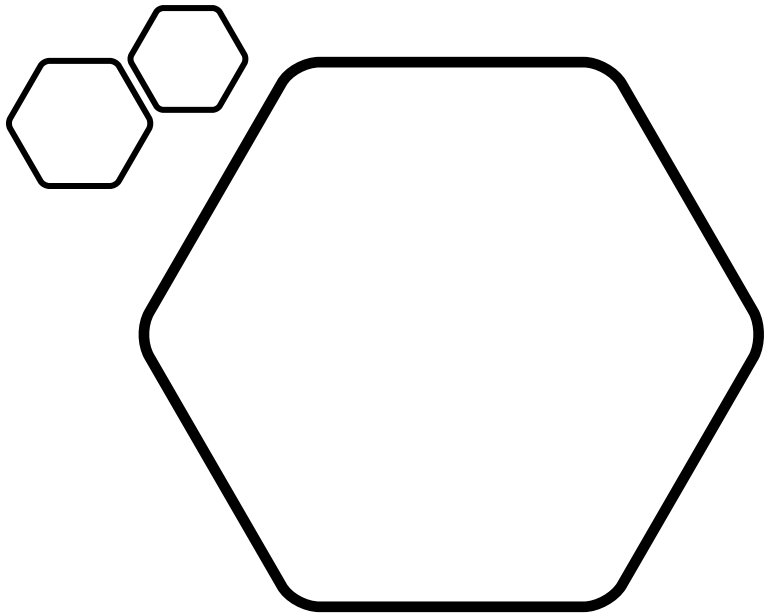
World Value Survey, 2008-2010



World Value Survey, 2017-2020

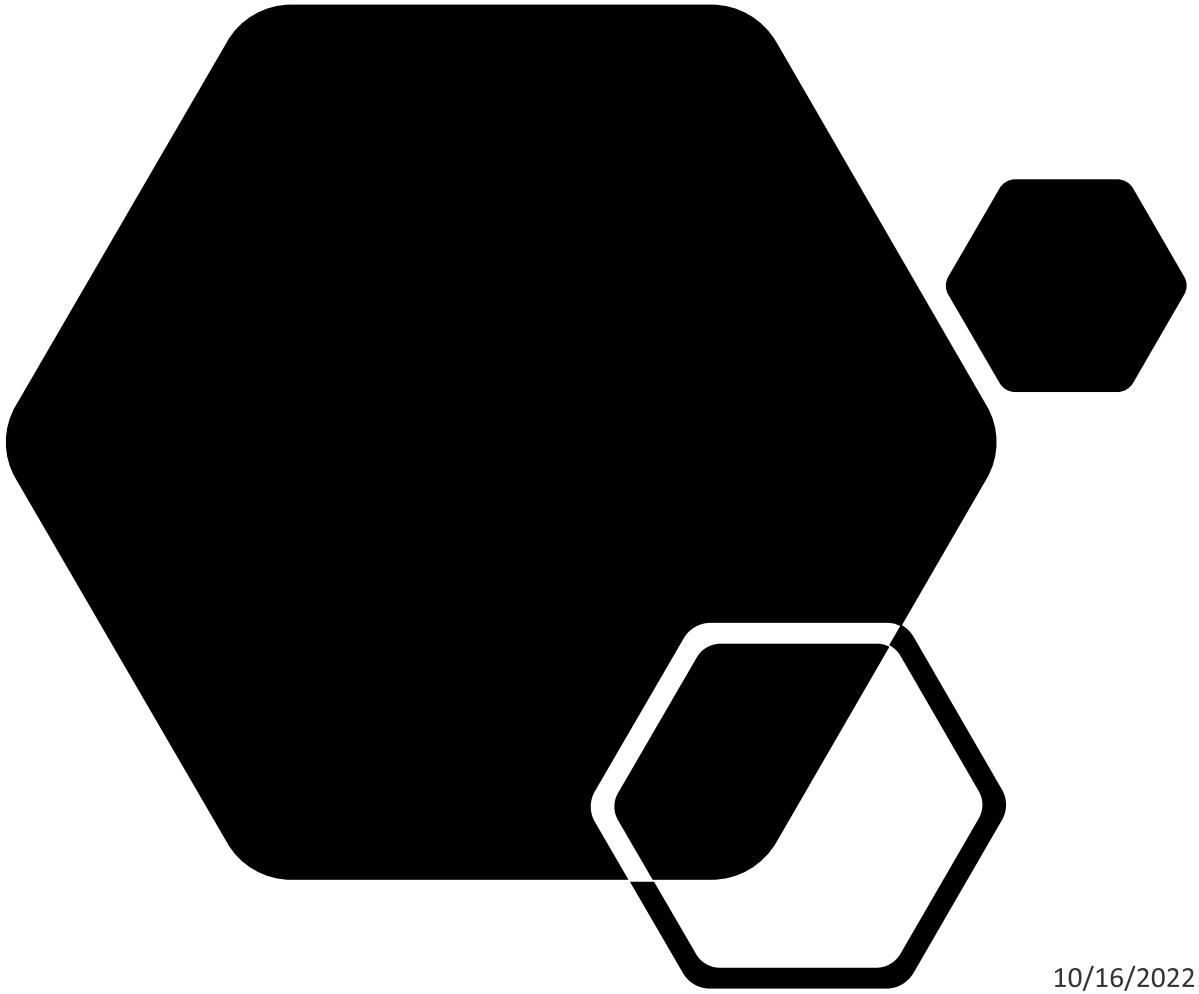
Concluzii

- Fenomenele sunt măsurate prin indicatori
- Procesul măsurat vs. procesul măsurării
- Două modele de măsurare
  - Modelul reflectiv: constructul cauzează indicatorii
  - Modelul formativ: indicatorii compun constructul
- Tendință centrală și variabilitate
- Erori de măsurare
  - Sistematice (validitate) vs. aleatorii (fidelitate)
  - Stilul de răspuns
  - Tipul de scală



Extra time

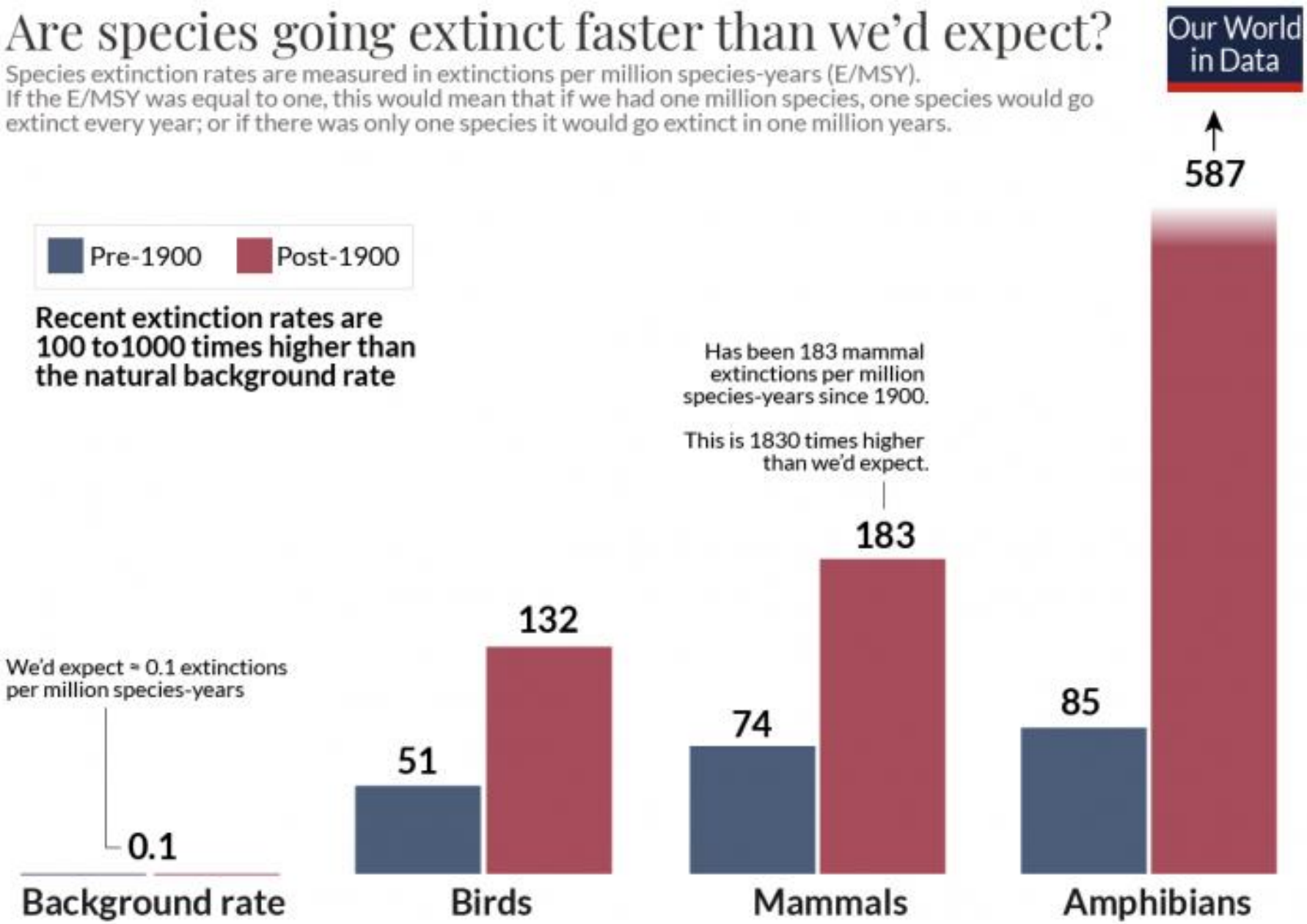
Subtitle



10/16/2022

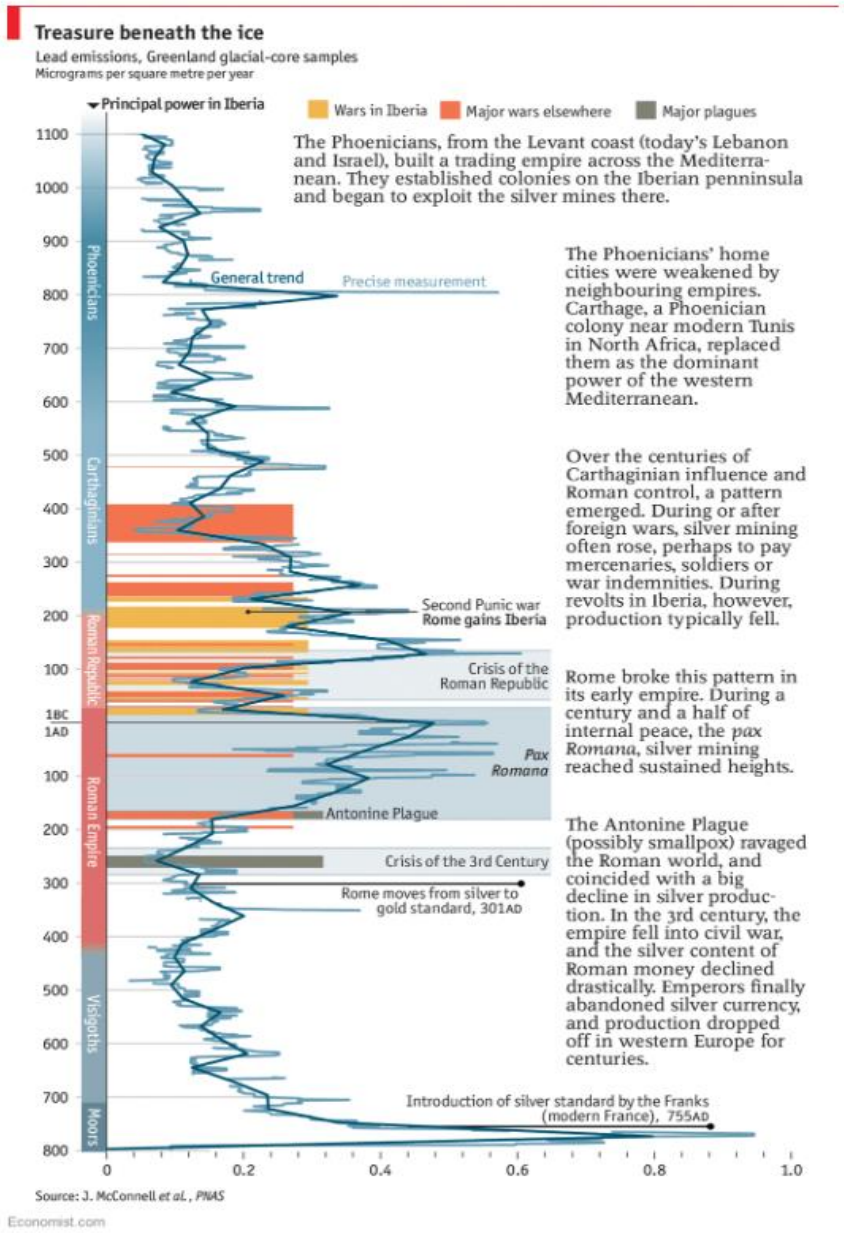
Diagnoza unei extincții masive

- O extincție masivă a speciilor este definită prin dispariția a peste 70% din specii într-o **perioadă scurtă**
  - Perioadă scurtă = 1-2 milioane de ani!
- Cum detectăm prezența sau absența unui proces de extincție masivă, cu date de câteva sute de ani?
  - „The Sixth Extinction” (vezi OWID – Extinctions)
  - Prin extrapolare
- Problema agregării datelor din epoci diferite
  - Scale temporale diferite: milioane vs. sute de ani
  - Similar cu datele din Keeling Curve
    - Măsurare directă a CO2 + inferențe din analiza ghețarilor


Are species going extinct faster than we'd expect?
Species extinction rates are measured in extinctions per million species-years (E/MSY).
If the E/MSY was equal to one, this would mean that if we had one million species, one species would go extinct every year; or if there was only one species it would go extinct in one million years.
Our World in Data
Pre-1900
Post-1900
Recent extinction rates are 100 to1000 times higher than the natural background rate
Has been 183 mammal extinctions per million species-years since 1900.
This is 1830 times higher than we'd expect.
We'd expect ≈ 0.1 extinctions per million species-years
0.1
51
132
74
183
85
587
Background rate
Birds
Mammals
Amphibians
Note: Species defined as 'probably exinct' by the IUCN are included as species extinctions.
Data Source: Pimm et al. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. Science.
OurWorldinData.org – Research and data to make progress against the world's largest problems.
Licensed under CC-BY by the author Hannah Ritchie.

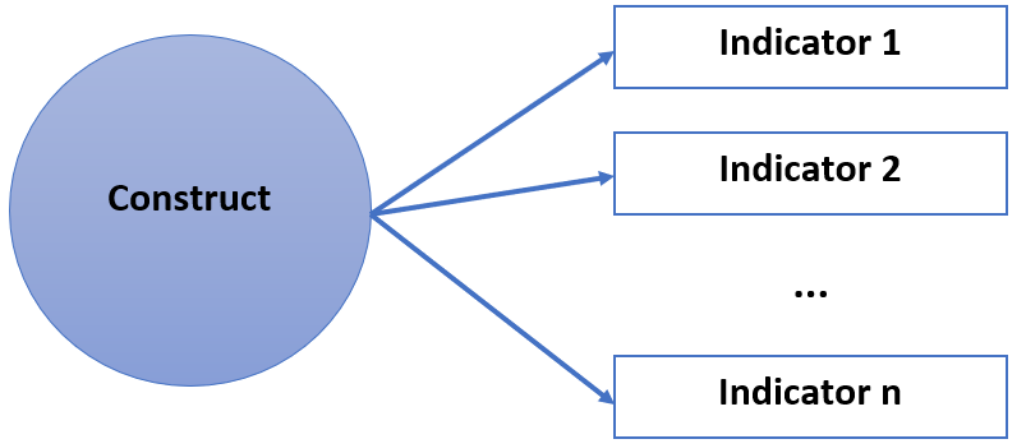
Sursa

Măsurarea trecutului prin prezent



Măsurarea circulației monetare în argint prin reziduurilor de plumb din gheața din Groenlanda

- Circulația monetară: fenomenul / constructul
- Reziduurile de plumb: indicator



Sursa