Curs 05 – Tehnici de analiză statistică

Analiza factorială

Analiza cluster

Analiza de regresie

Analiza de rețea

Serii de timp

10/26/2022

Structura cursului

1. Why?
2. Cauzalitate
3. Măsurare
4. Modelare și eșantionare
5. Tehnici de analiză
   - Analiza factorială
   - Analiza cluster
   - Analiza de regresie
   - Analiza de rețea
   - Serii de timp
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

Obiective ale tehnicilor de analiză

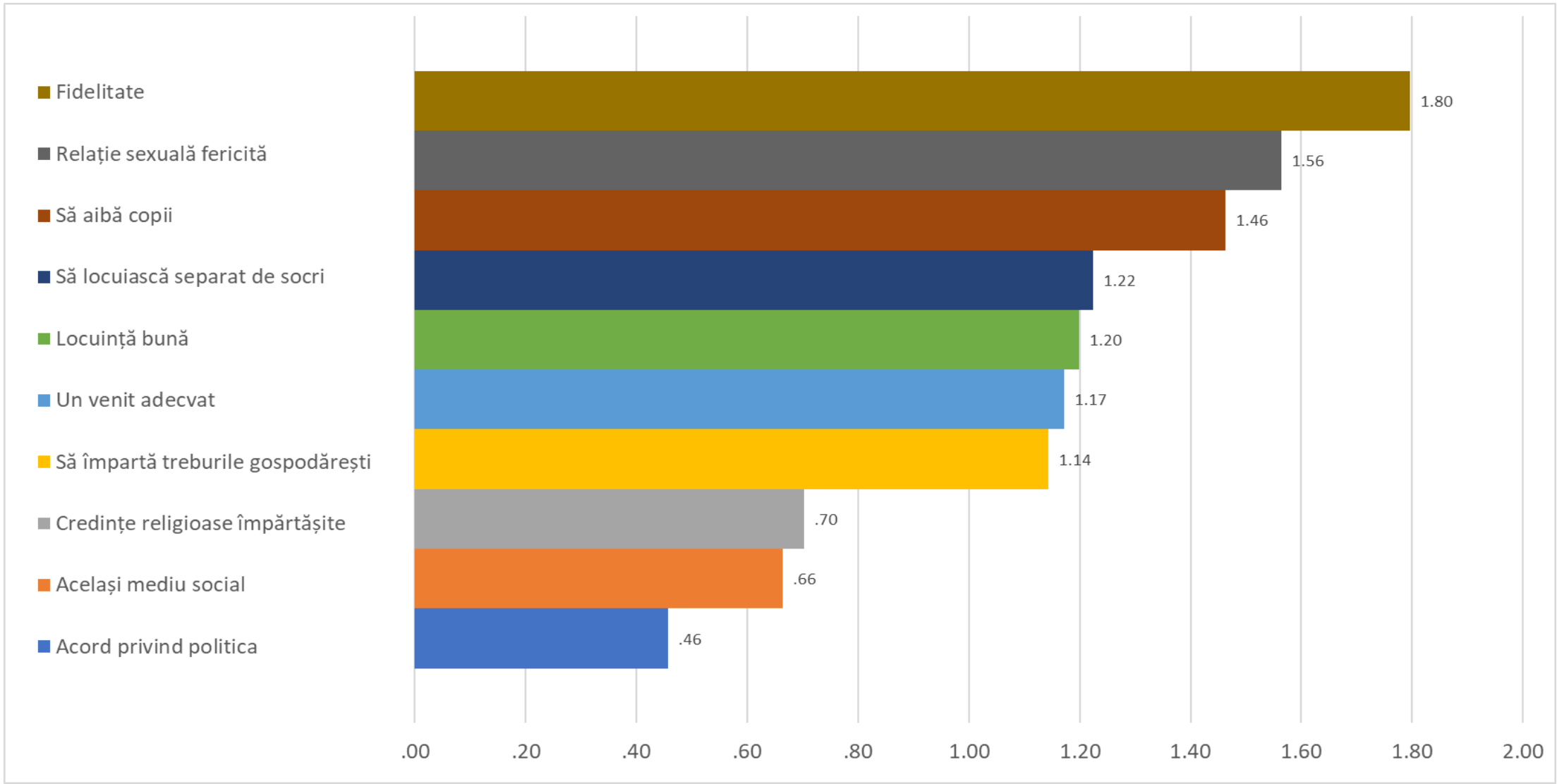
- Explorarea datelor
  - Frecvențe, medii și corelații
- Măsurare: analiza factorială / reducerea dimensionalității
  - Identificarea dimensiunilor personalității pe baza acțiunilor și preferințelor
- Clasificare: analiza cluster
  - Clasificarea indivizilor în tipuri: „Spune-mi cu cin’ te-nsoțești, ca să-ți spun cine ești”
- Explicare:
  - Analiza de regresie:
    - Explicarea acțiunilor prin factori externi – educație, vârstă, venit etc.
  - Analiza de rețea
    - Explicații prin proximitate, contagiune: „Cin’ se-aseamănă, se-adună”
- Extrapolare: seriile de timp
  - Explicarea acțiunilor prin factori externi și patternuri temporale
    - Tendințe inerțiale, sezoniere, variații aleatorii

Studiu de caz

Aici sunt câteva aspecte despre care unii oameni cred că sunt importante pentru o căsătorie de succes. Te rog, pentru fiecare, spune-mi cum crezi că este, pentru succesul unei căsătorii: (2=foarte important, 1=destul de important, 0=nu prea important)

- Fidelitatea
- Partenerii să fie din același mediu social
- Să aibă aceeași religie
- Să aibă o locuință bună
- Un venit adecvat
- Să fie de acord în chestiunile politice
- Să trăiască separat de socri
- Să aibă o relație sexuală fericită
- Să împartă treburile domestice
- Să aibă copii

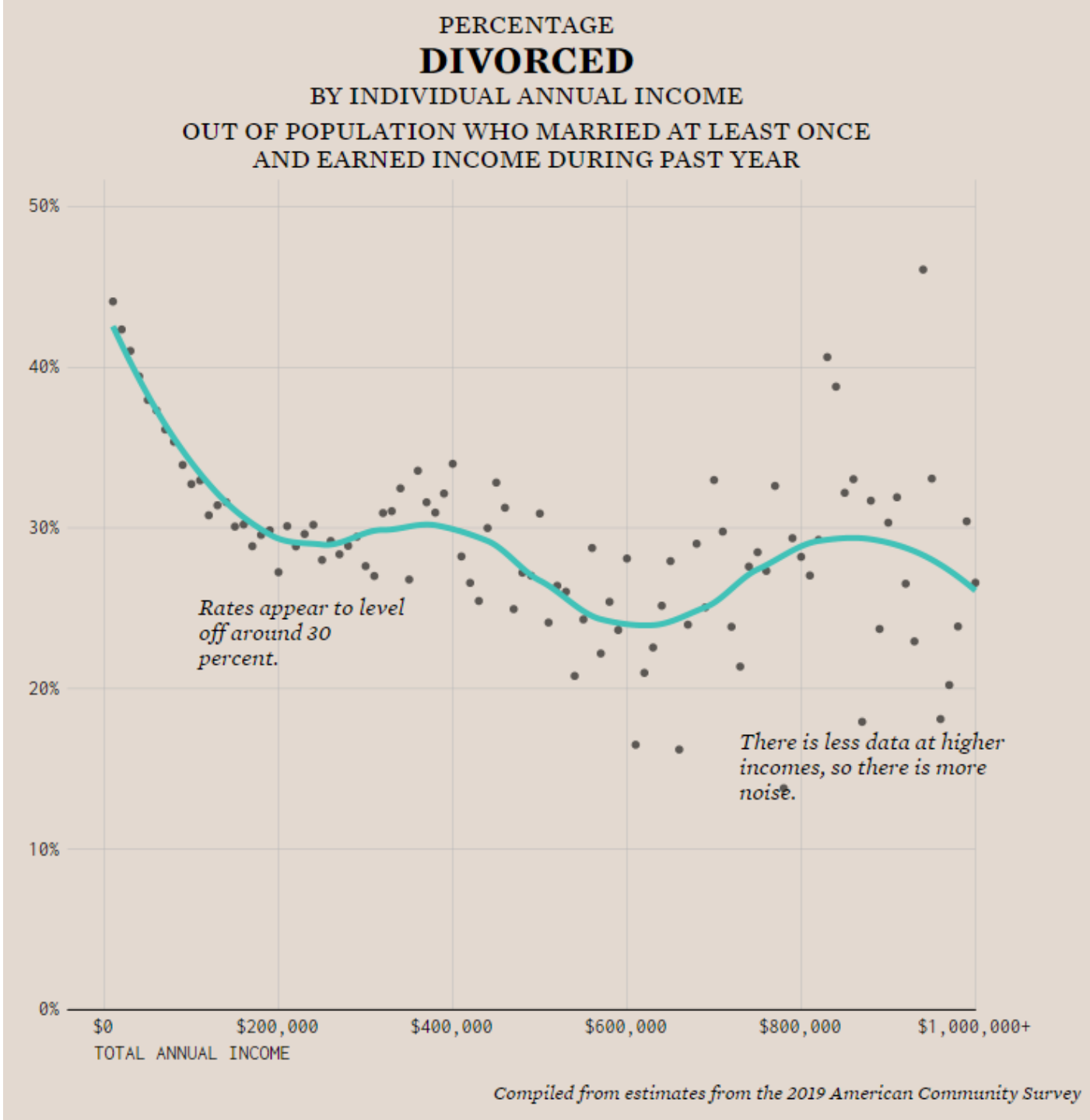
Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia



Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Percepție vs. realitate

- O relație clară între venit și divorțialitate



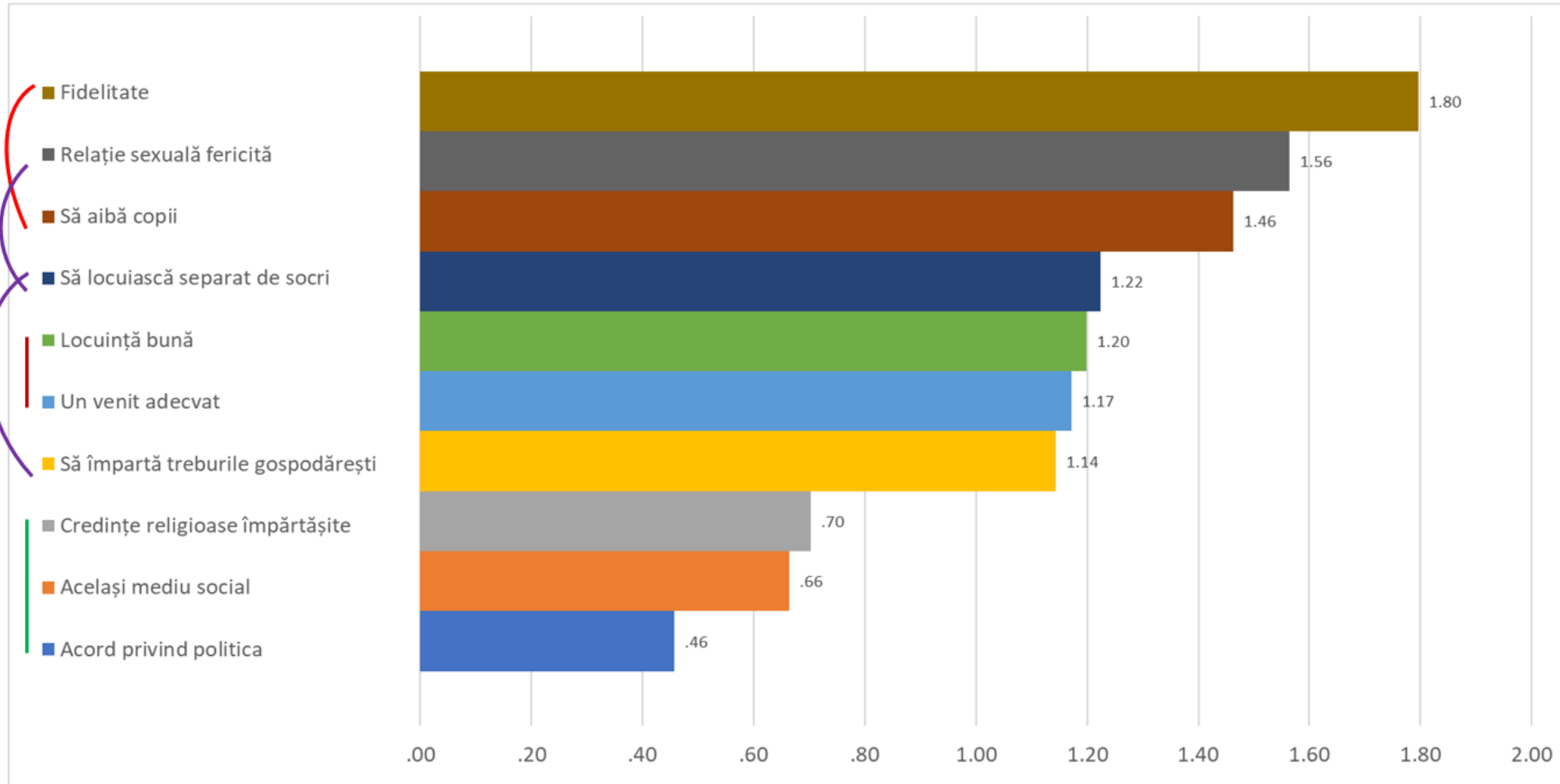
Sursa

Tabelul de corelații

	Fidelitate	Copii	Venit	Casă	Același mediu	Aceeași religie	Acord în politică	Locuiesc singuri	Relație sexuală	Împart treburile
Fidelitate	1	.223	.079	.088	.058	.160	.028	-0.021	.041	.085
Copii	.223	1	.138	.177	.102	.192	.058	.034	.111	.175
Venit	.079	.138	1	.454	.298	.146	.143	.090	.134	.096
Casă	.088	.177	.454	1	.252	.191	.186	.116	.144	.185
Același mediu	.058	.102	.298	.252	1	.355	.301	.079	.027	.063
Aceeași religie	.160	.192	.146	.191	.355	1	.316	-.013	-.027	.074
Acord în politică	.028	.058	.143	.186	.301	.316	1	.065	.051	.122
Locuiesc singuri	-.021	.034	.090	.116	.079	-.013*	.065	1	.247	.137
Relație sexuală	.041	.111	.134	.144	.027	-.027	.051	.247	1	.233
Împart treburile	.085	.175	.096	.185	.063	.074	.122	.137	.233	1

Coeficienți de corelație Bravais-Pearson

Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

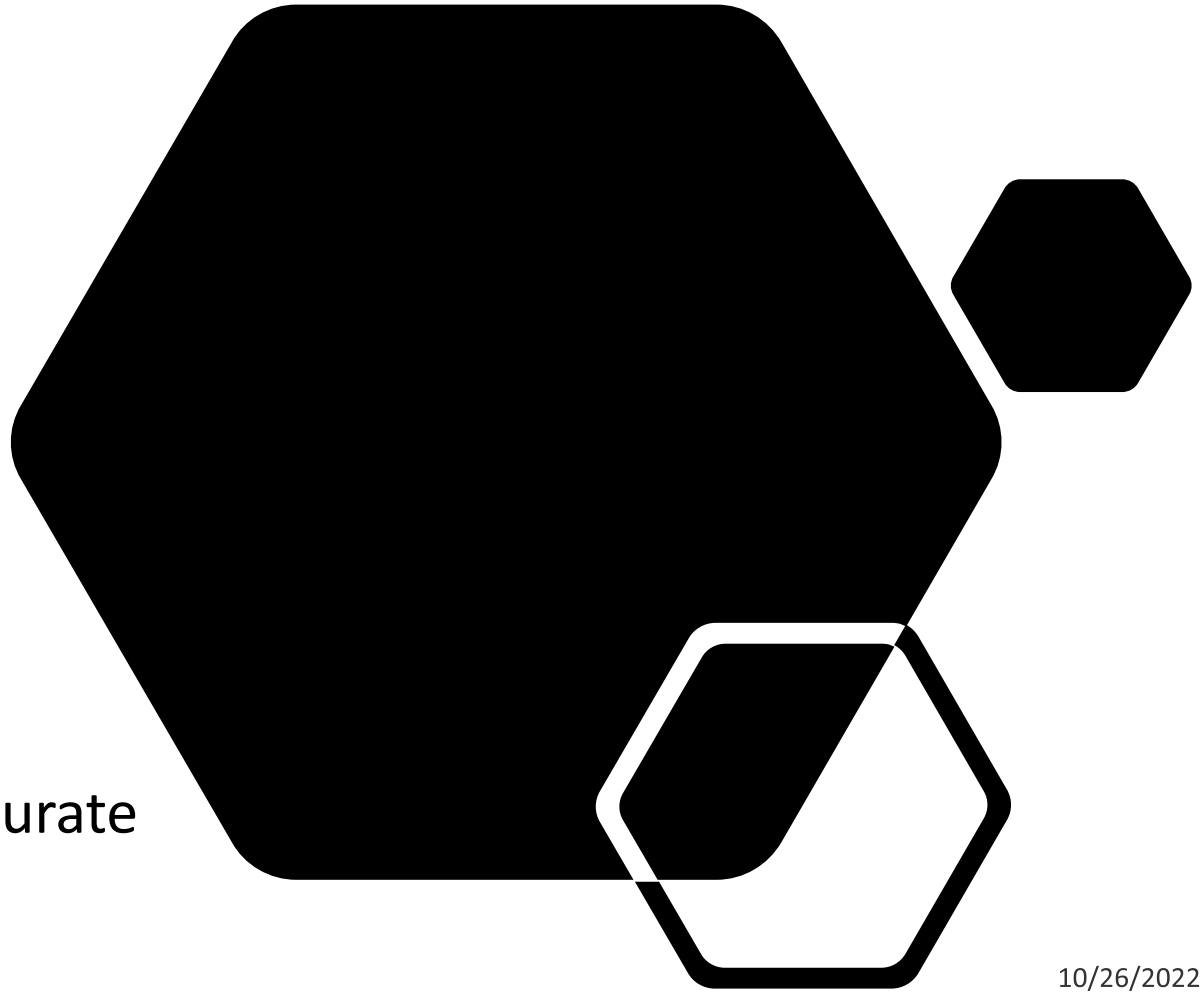


Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Analiza factorială

Reducerea dimensionalității

Trecerea de la indicatori la constructele măsurate (dimensiuni)

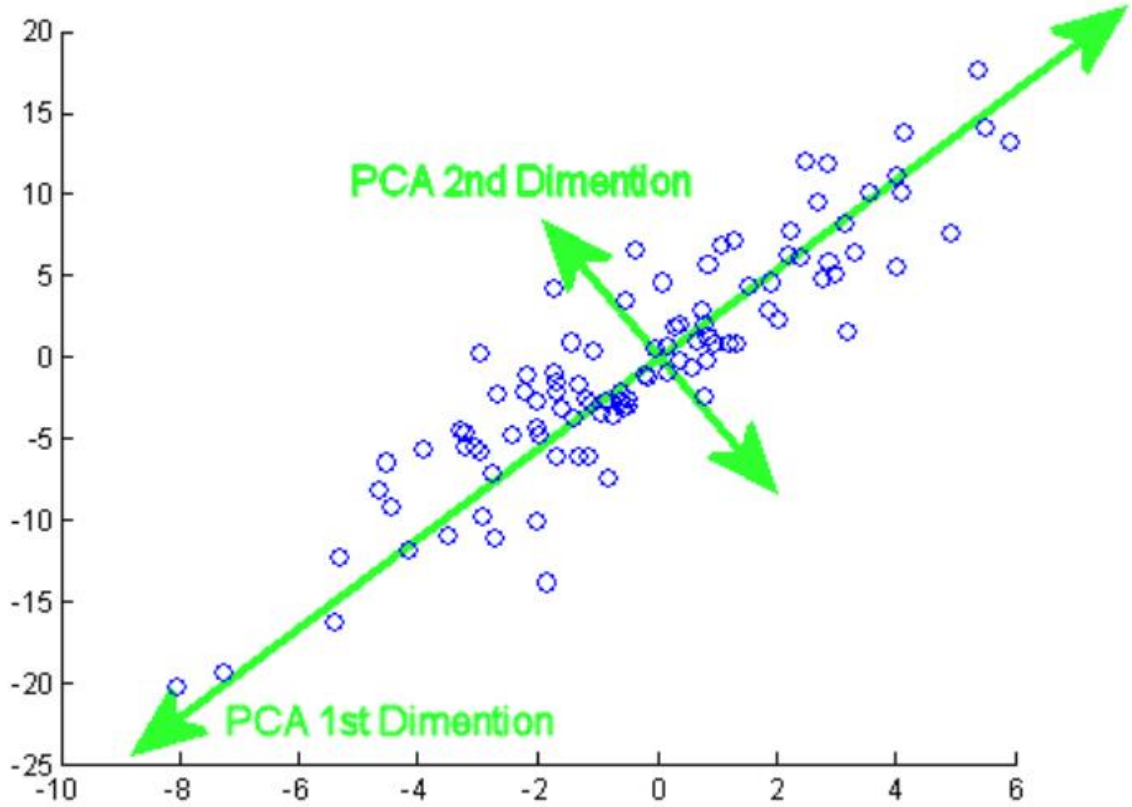


Analiza factorială

- Estimează...
  - **Conceptele măsurate** de o diversitate de indicatori corelați
  - **Forțele** care se află în spatele unei diversități de simptome corelate
    - Acestea sunt factori / dimensiuni
    - Interpretăm semnificația factorilor în funcție de simptomele asociate
- O analiză factorială **generează n variabile**
  - Fiecare factor e o nouă variabilă, cu valori pentru toate cazurile analizate
    - „Scorul factorial” are valori continue, reale
    - Extragem mai puțini factori decât indicatorii inițiali
- Fiecare individ analizat primește valori estimate pentru fiecare factor extras

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
- Exemplu grafic
  - Doi factori / două dimensiuni
  - Dimensiunea 1 este mai importantă



Sursă

Patru factori identificați pentru 10 itemi: cum îi interpretăm?

Fiecare factor este o nouă variabilă

Fiecare respondent are o valoare pentru fiecare factor

Similaritate

Intimitate

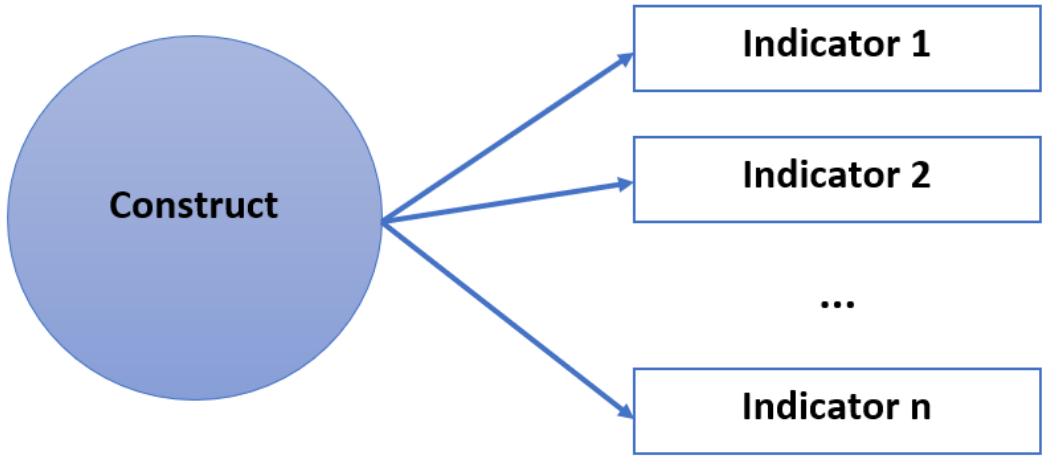
Familia tradițională

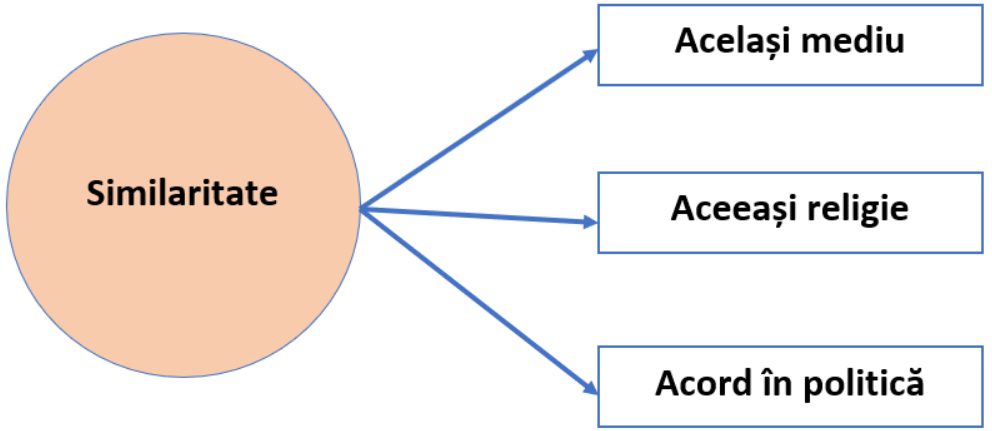
Banii NU contează

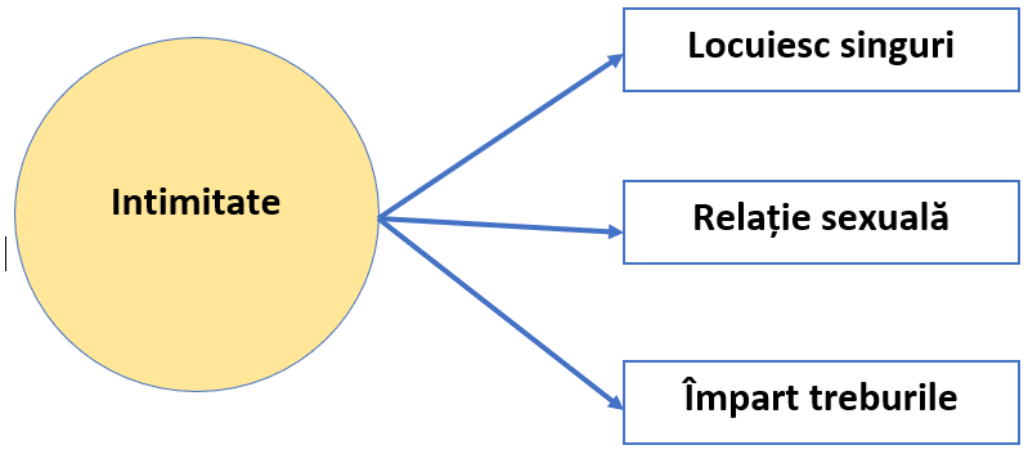
	Factor			
	1	2	3	4
Fidelitate	-.002	-.070	.766	.005
Copii	.016	.118	.701	-.100
Venit	-.022	-.024	.003	-.875
Locuință bună	.038	.081	.077	-.771
Același mediu	.619	-.058	-.080	-.316
Aceeași religie	.737	-.121	.244	.028
Acord în politică	.792	.140	-.113	.078
Locuiesc singuri	.053	.680	-.224	-.035
Relație sexuală	-.122	.726	.053	-.087
Împart treburile	.095	.606	.270	.068

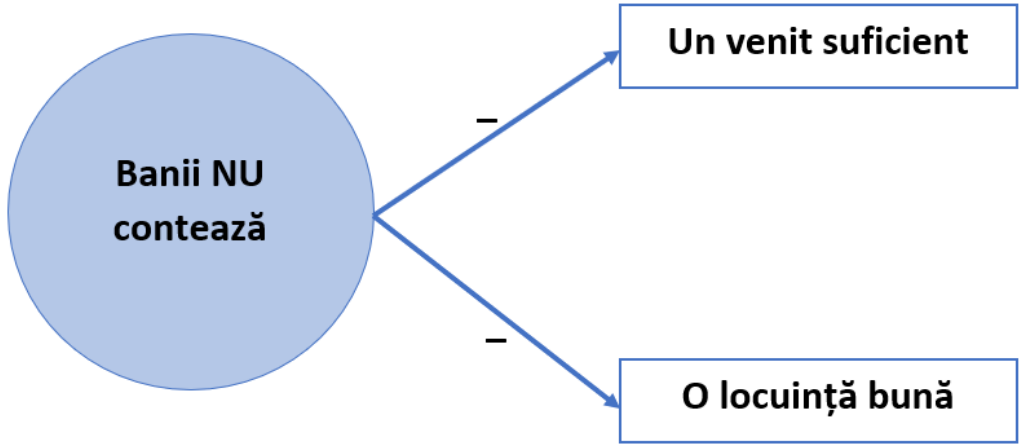
Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Modelul reflectiv de măsurare



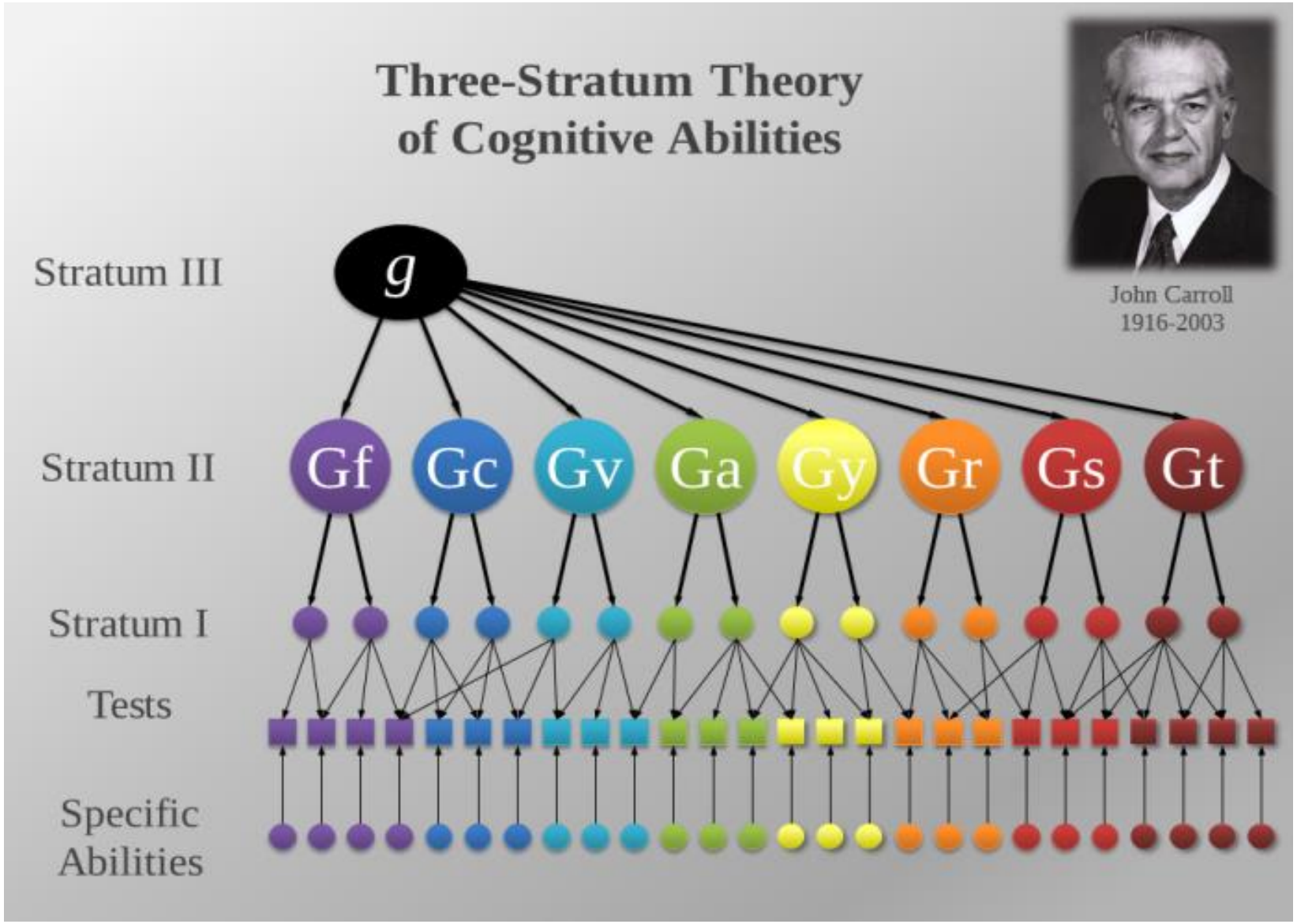






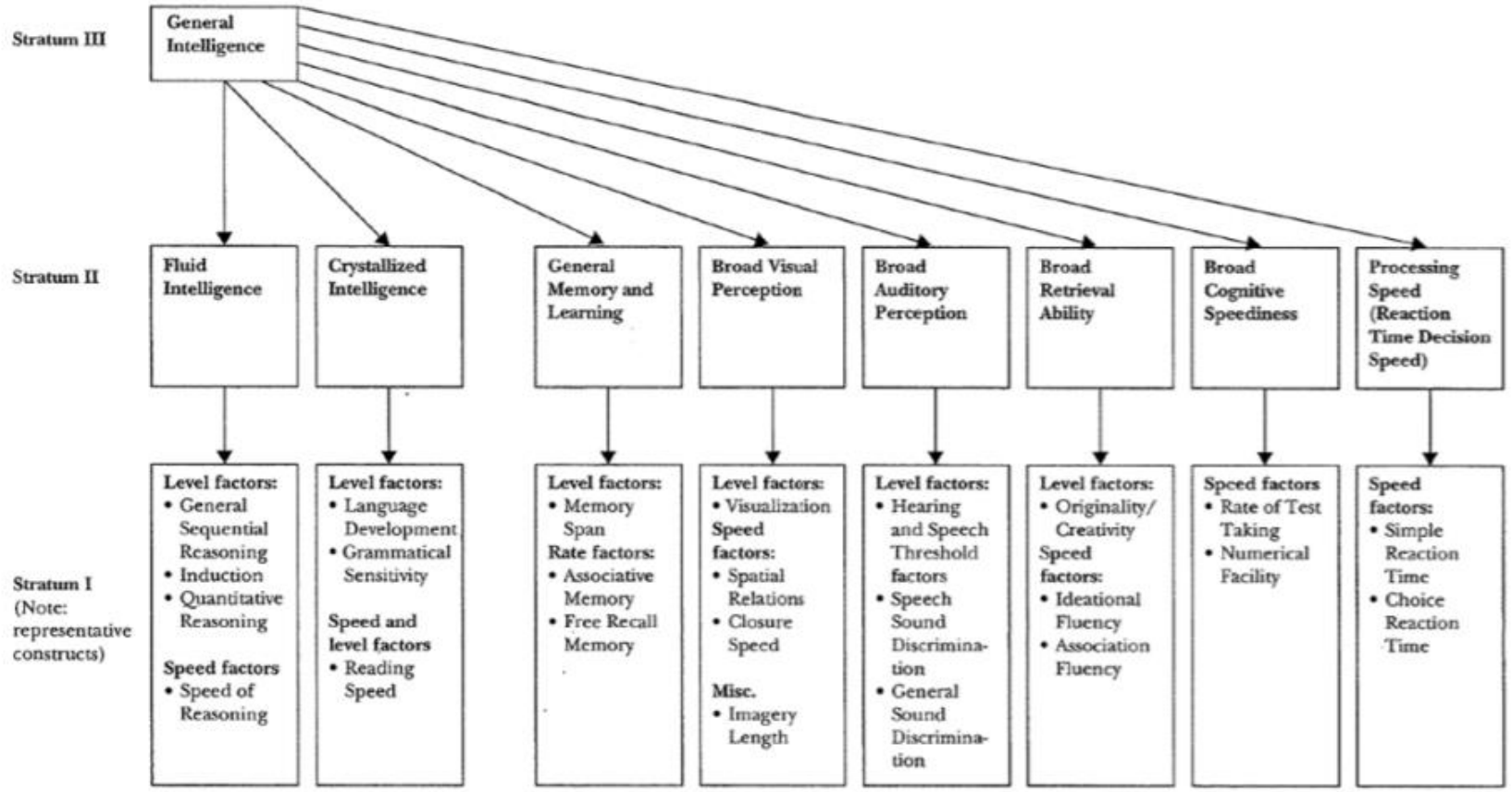
Măsurarea inteligenței & IQ

- 3 straturi



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

Carroll's Three-Stratum Theory of Cognitive Ability



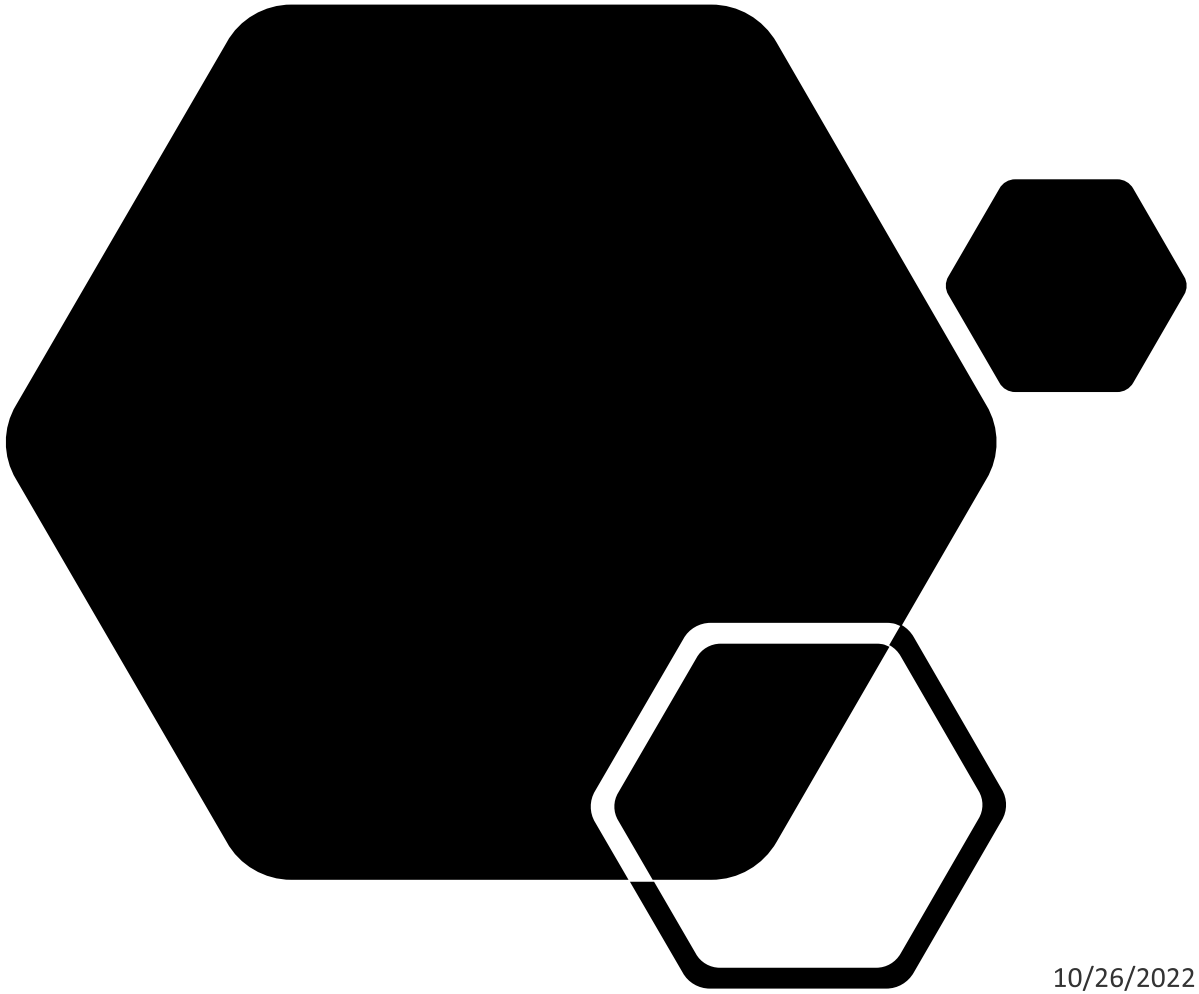
Adapted with permission from Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press. (p. 626).

6

McDaniel, 2018

Analiza cluster

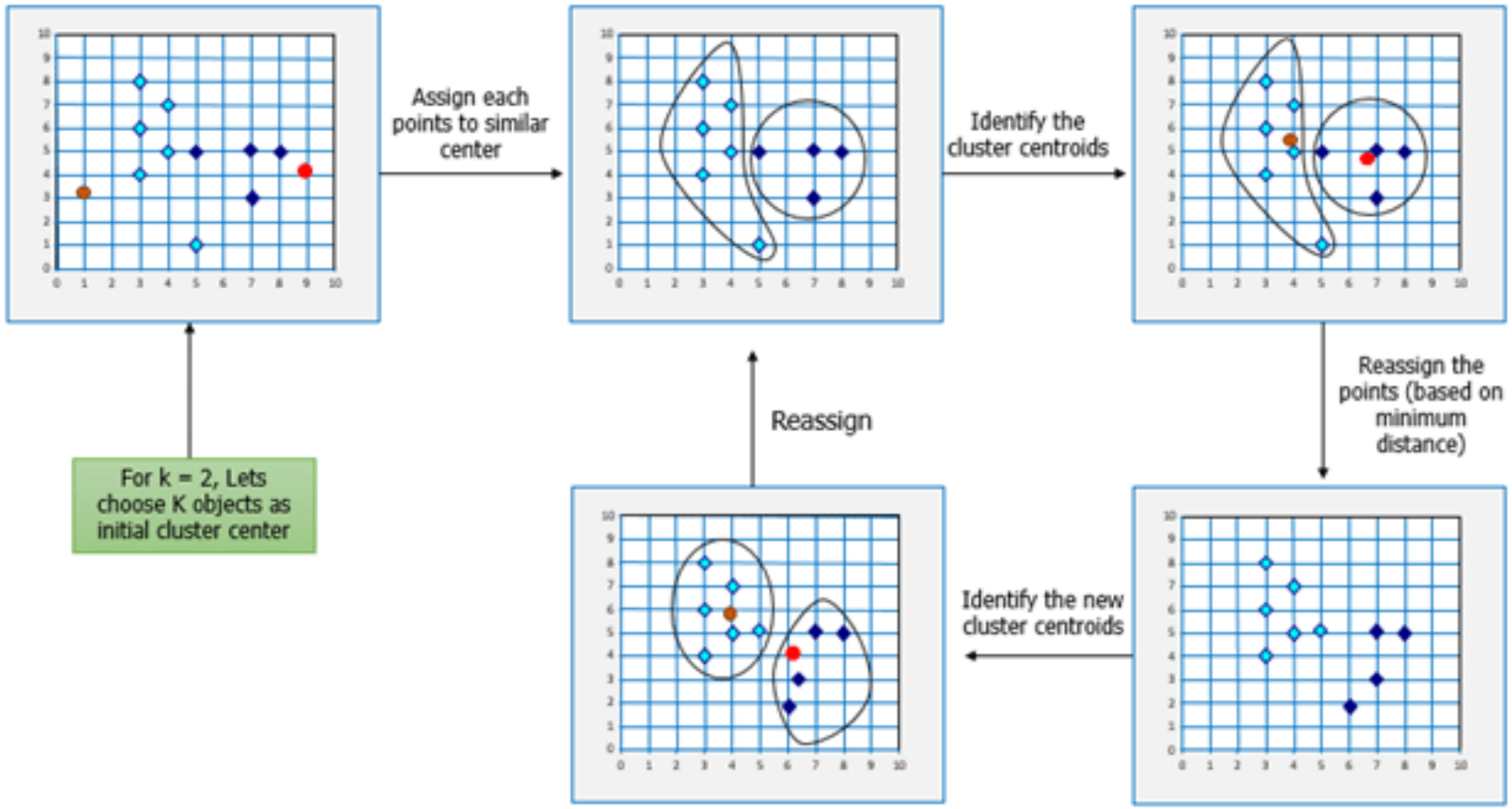
Identificarea unor tipologii



Analiza cluster

- Identifică...
  - **Tipurile** de indivizi în funcție de o serie de criterii
  - Cluster = tip
  - Interpretăm semnificația tipurilor în funcție de profilul lor pe criteriile analizate
  - Putem explora clasificări cu mai multe sau mai puține tipuri
- O analiză cluster generează **o singură variabilă cu n valori**
  - Valori naturale, pozitive (de la 1 la n)
  - Fiecare valoare este un tip
- Fiecare caz este inclus într-un tip și numai unul

Analiza cluster K-means



Cuplu cu copii, familie extinsă

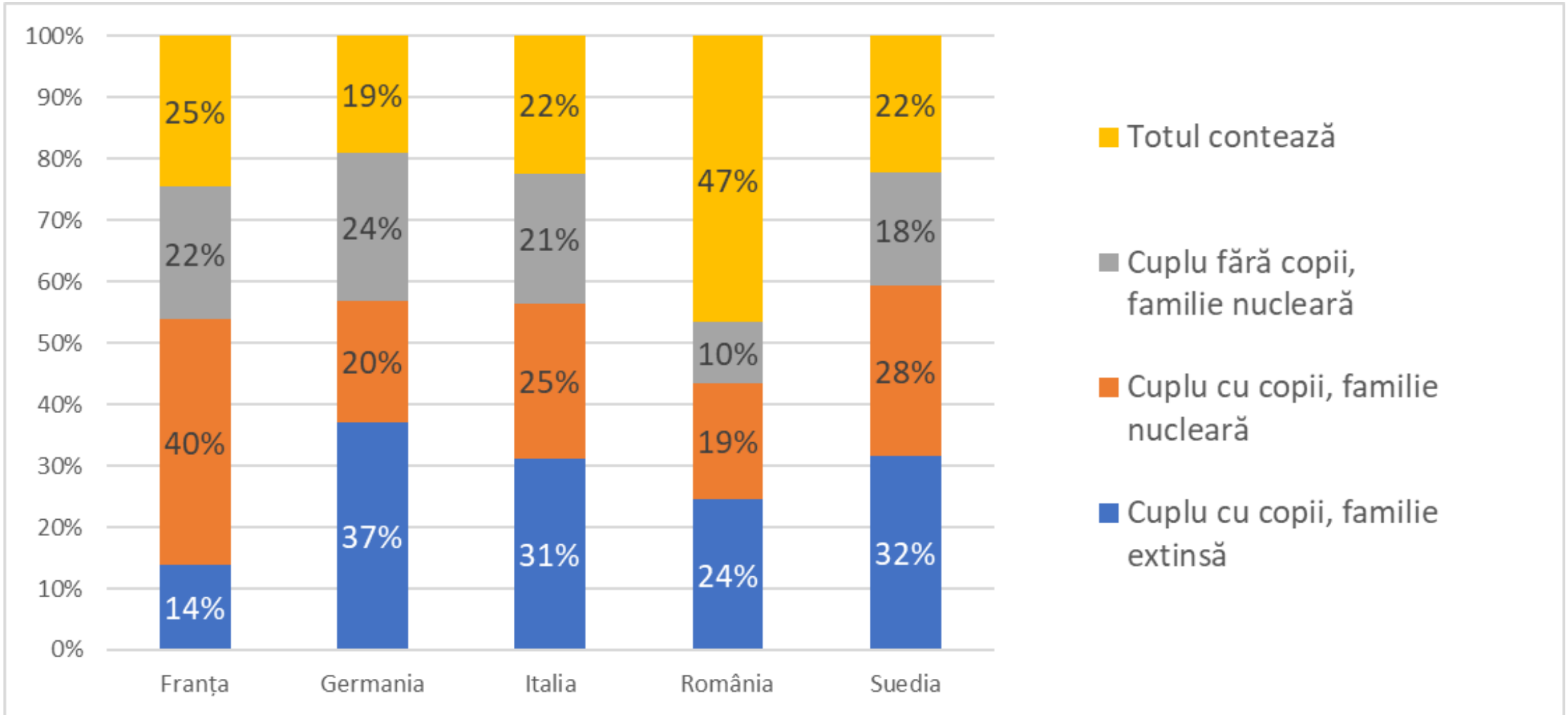
Cuplu cu copii, familie nucleară

Cuplu fără copii, familie nucleară

Totul contează

	Cluster			
	1	2	3	4
Fidelitate	1.81	1.86	1.58	1.91
Copii	1.42	1.87	.64	1.77
Relație sexuală	1.25	1.79	1.58	1.69
Locuință bună	.92	1.25	1.00	1.64
Venit	.90	1.15	1.05	1.60
Împart treburile	.85	1.46	.86	1.40
Aceeași religie	.63	.31	.30	1.49
Același mediu	.41	.31	.55	1.41
Locuiesc singuri	.32	1.71	1.65	1.41
Acord în politică	.30	.25	.33	.96

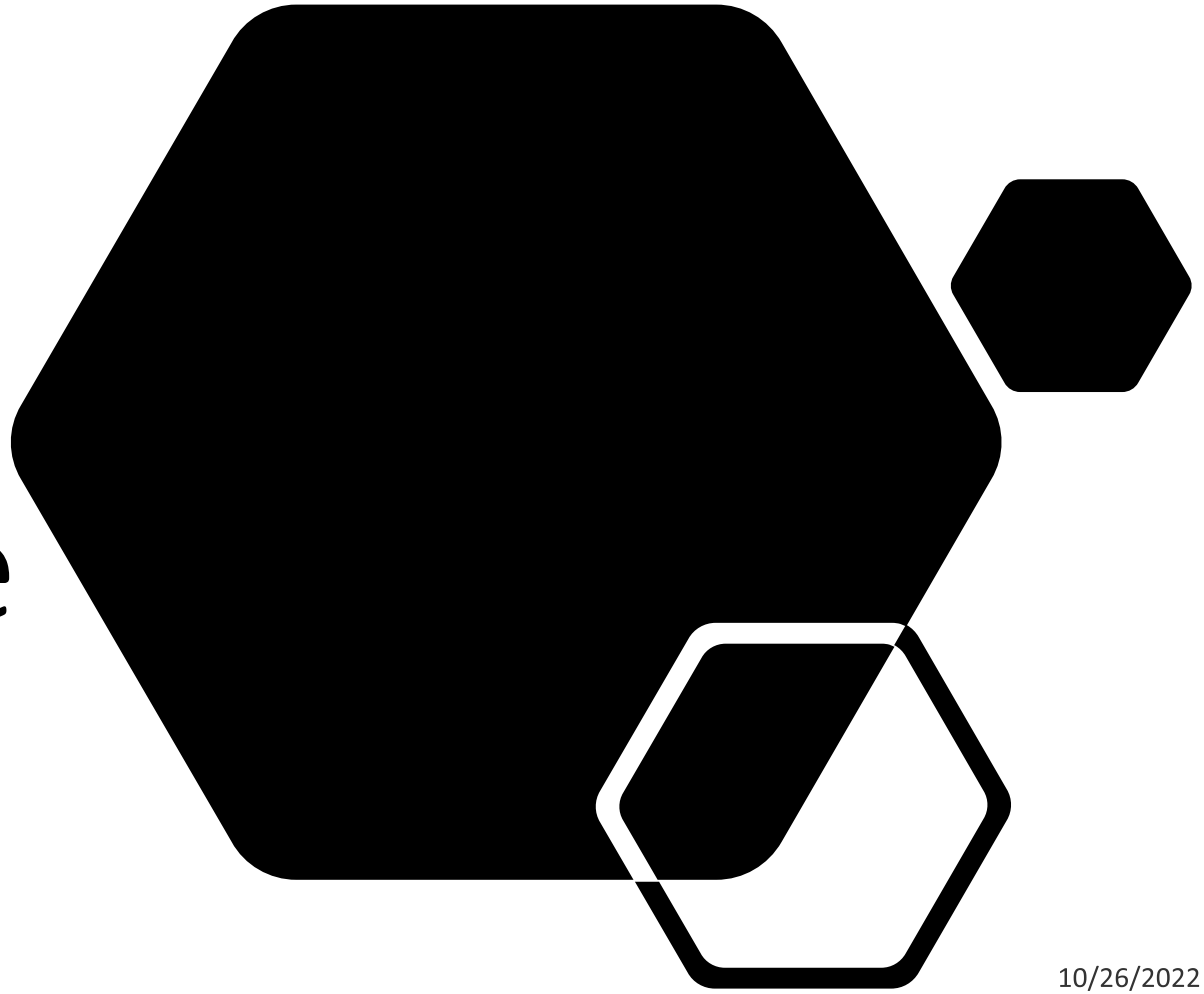
Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia



Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Analiza de regresie

Explicații prin factori externi



Analiza de regresie

- Modelează variația unui efect ca funcție a mai multor predictori
  - Funcție liniară, polinomială, exponențială etc
  - Regresie liniară: $Y = a + b_1*x_1 + b_2*x_2 +...+b_n*x_n$
- Corelație: relații bivariate
- Regresie: modele multivariate
  - Relațiile dintre Y și cei n predictori sunt estimate simultan
  - Dacă se adaugă / scoate un X, se modifică toate estimările
  - Coeficientul b_i sau $beta_i$ arată relevanța predictivă a lui x_i când restul de x sunt „ținuți sub control”, adică sunt constanți
  - Coeficienții beta lucrează cu unități de măsură standardizate (abateri standard) și devin comparabili

Regresia ca bază a predicției

- O regresie estimată permite anticiparea unor valori Y necunoscute pentru care știm valorile X_i
  - Cu cât crește proporția de premii Nobel în funcție de consumul de ciocolată, vin și lapte

Analiza de regresie: teoriile căsătoriei de succes

	Similaritate R Square: 23%		**Intimitate R Square: 21%**		**Familia tradițională R Square: 17%**		**Banii nu contează R Square: 15%**	
	Beta	Sig.	**Beta**	Sig.	**Beta**	Sig.	**Beta**	Sig.
Sex - Feminin	**.065**	.000	**-.012**	.137	**.081**	.000	**.014**	.077
Vârstă	**.220**	.000	**-.195**	.000	**.105**	.000	**-.143**	.000
Mărimea localității	**-.003**	.677	**.060**	.000	**-.102**	.000	**.026**	.001

Când controlăm genul și mărimea localității, vârsta este cel mai puternic predictor pentru teoriile indivizilor privind succesul în căsătorie

- Relații pozitive cu teoriile similarității, familiei tradiționale
- Relații negative cu teoriile intimității și „banii nu contează”

Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

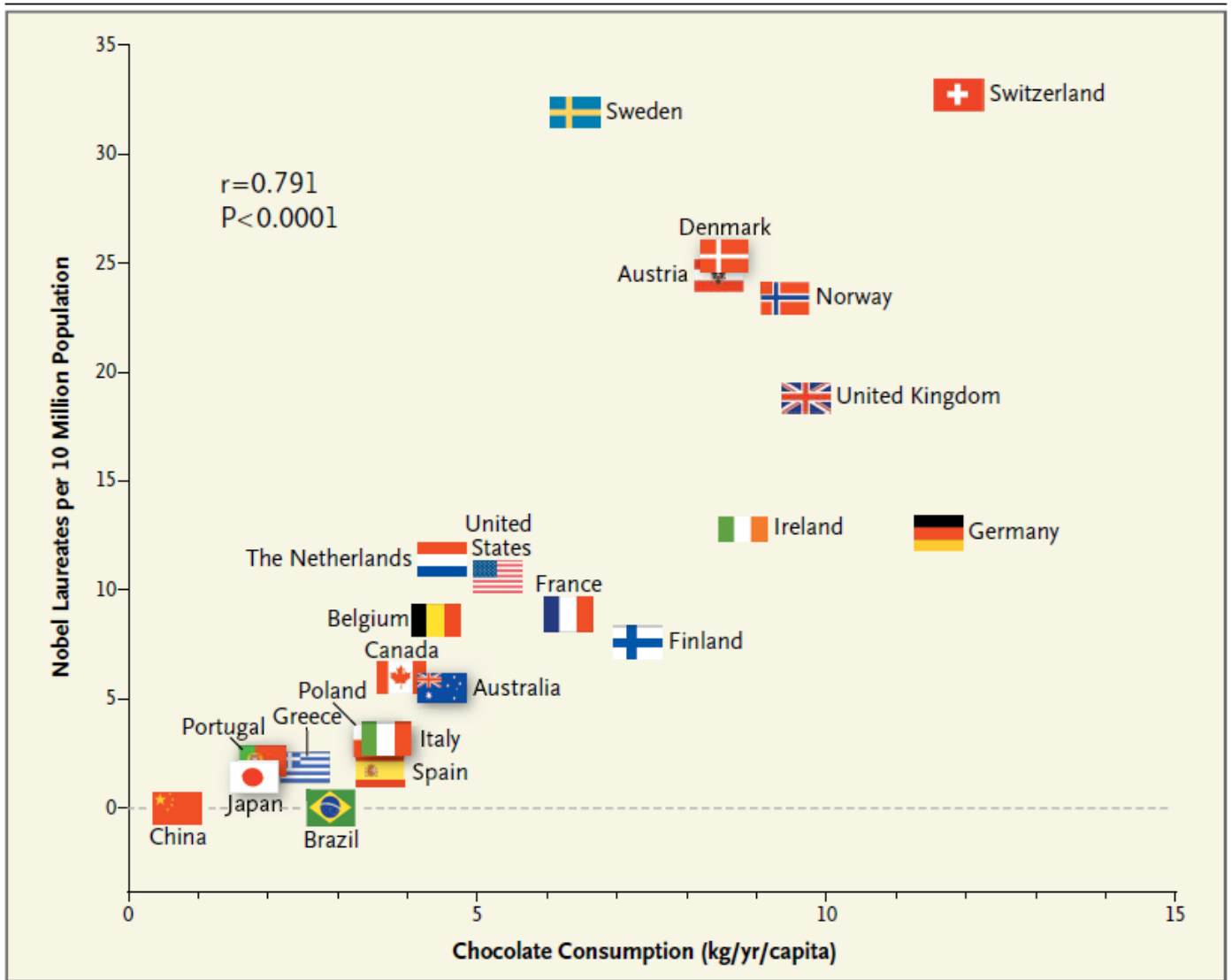
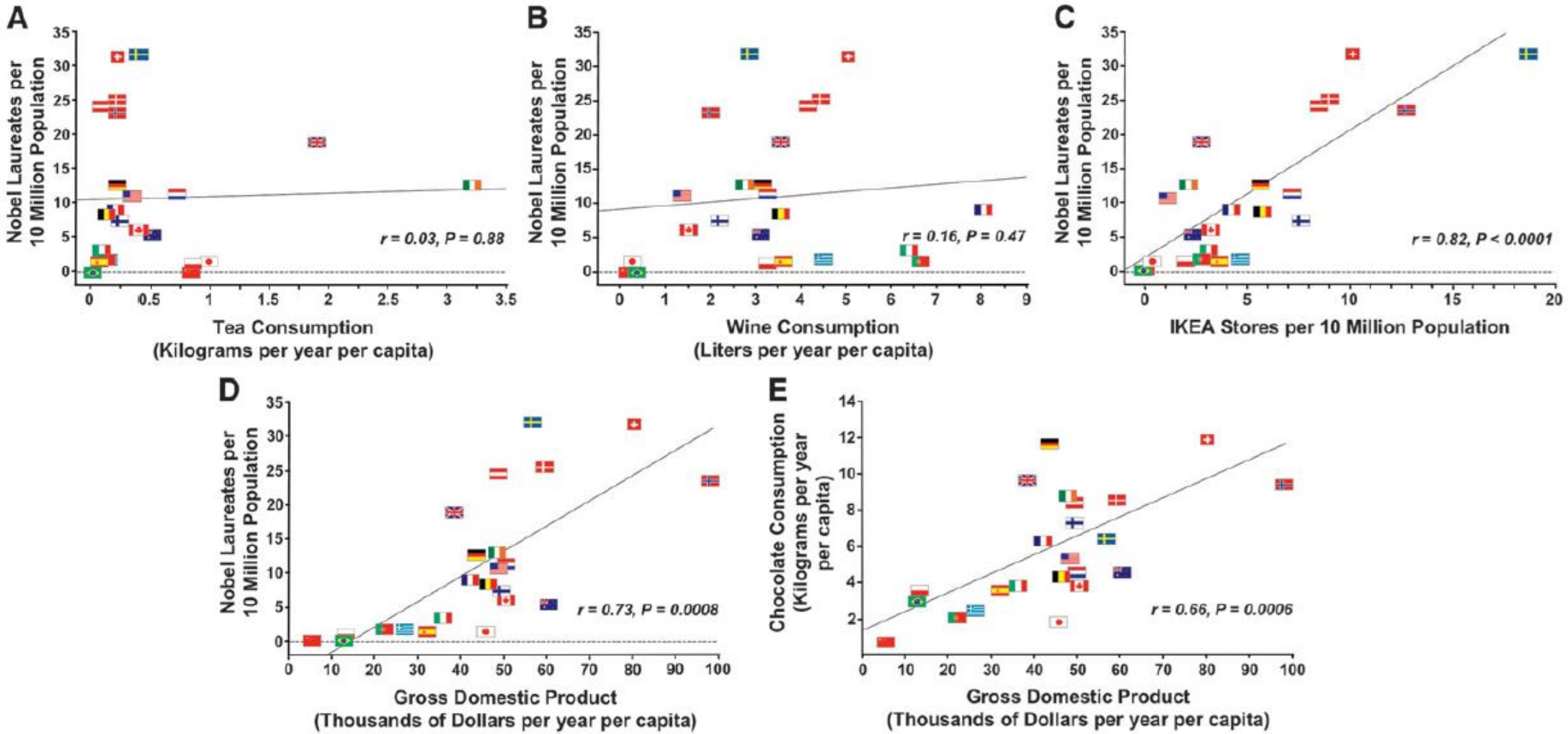
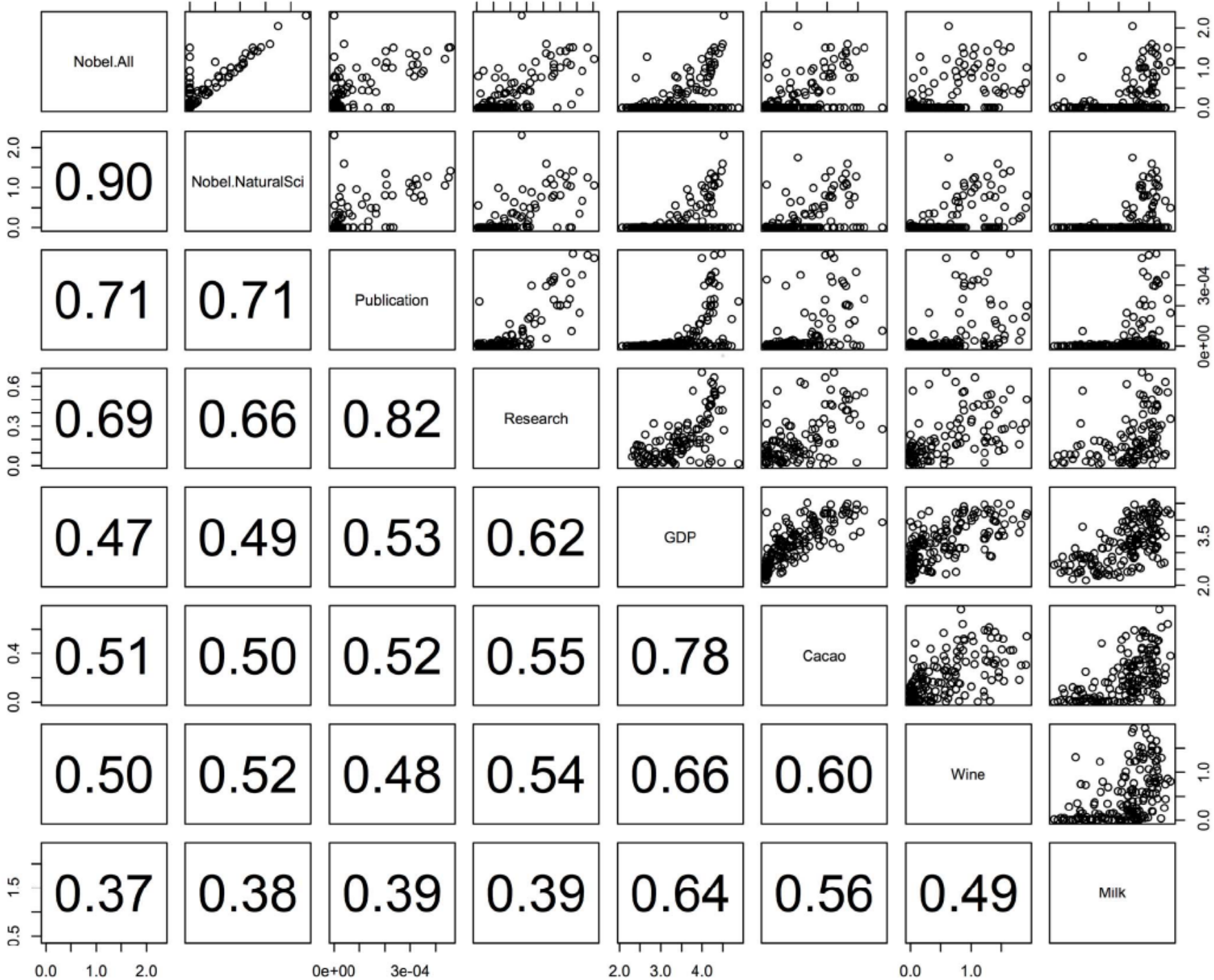


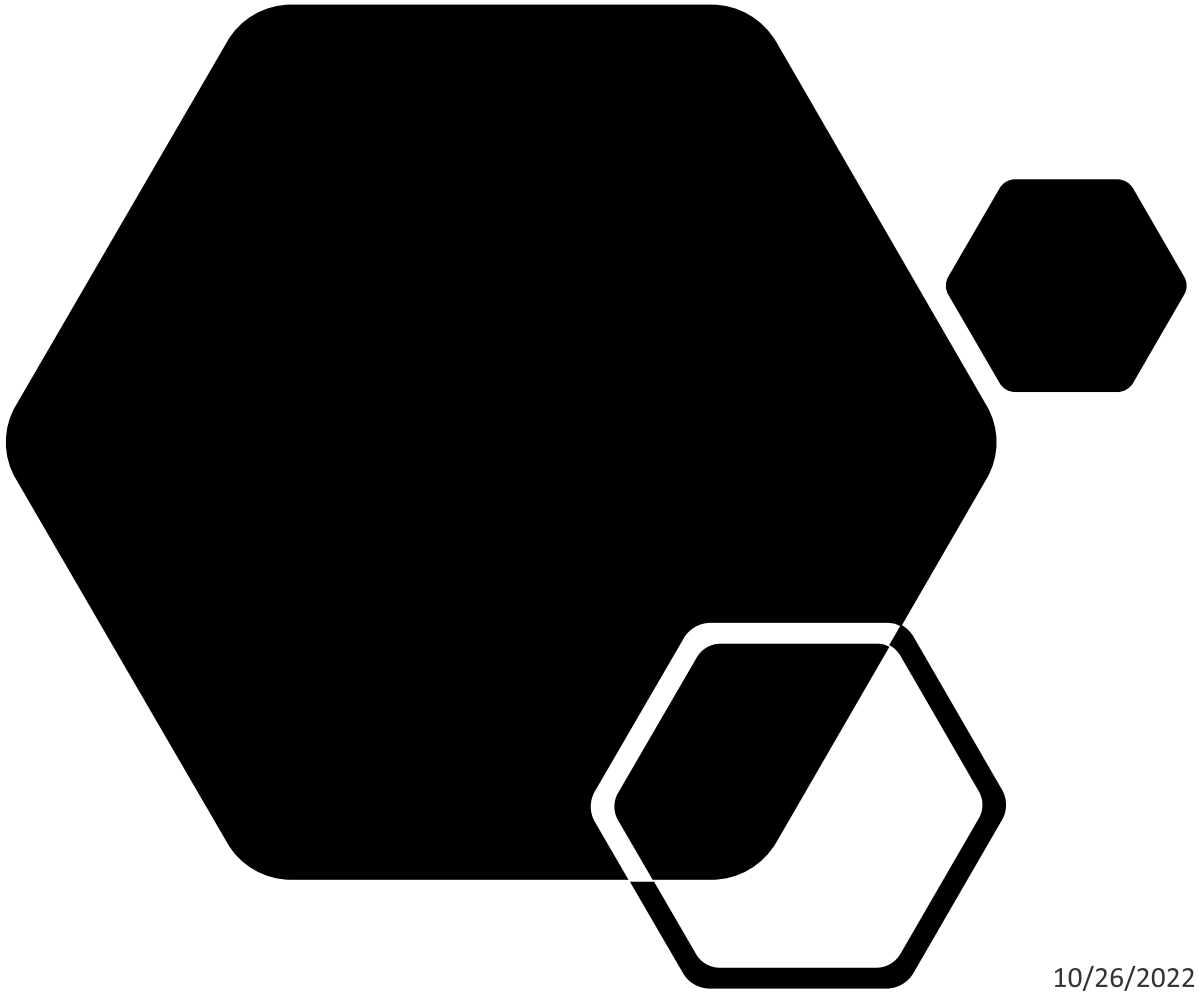
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.


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
0.90
Publication
0.71
0.71
Research
0.69
0.66
0.82
GDP
0.47
0.49
0.53
0.62
Cacao
0.51
0.50
0.52
0.55
0.78
Wine
0.50
0.52
0.48
0.54
0.66
0.60
Milk
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
0.5
1.5
0.0
0.4
2.0
3.5
0.0
0.3
0.6
0e+00
3e-04
0.0
1.0
2.0

Analiza de rețea

Explicații prin proximitate, contagiune



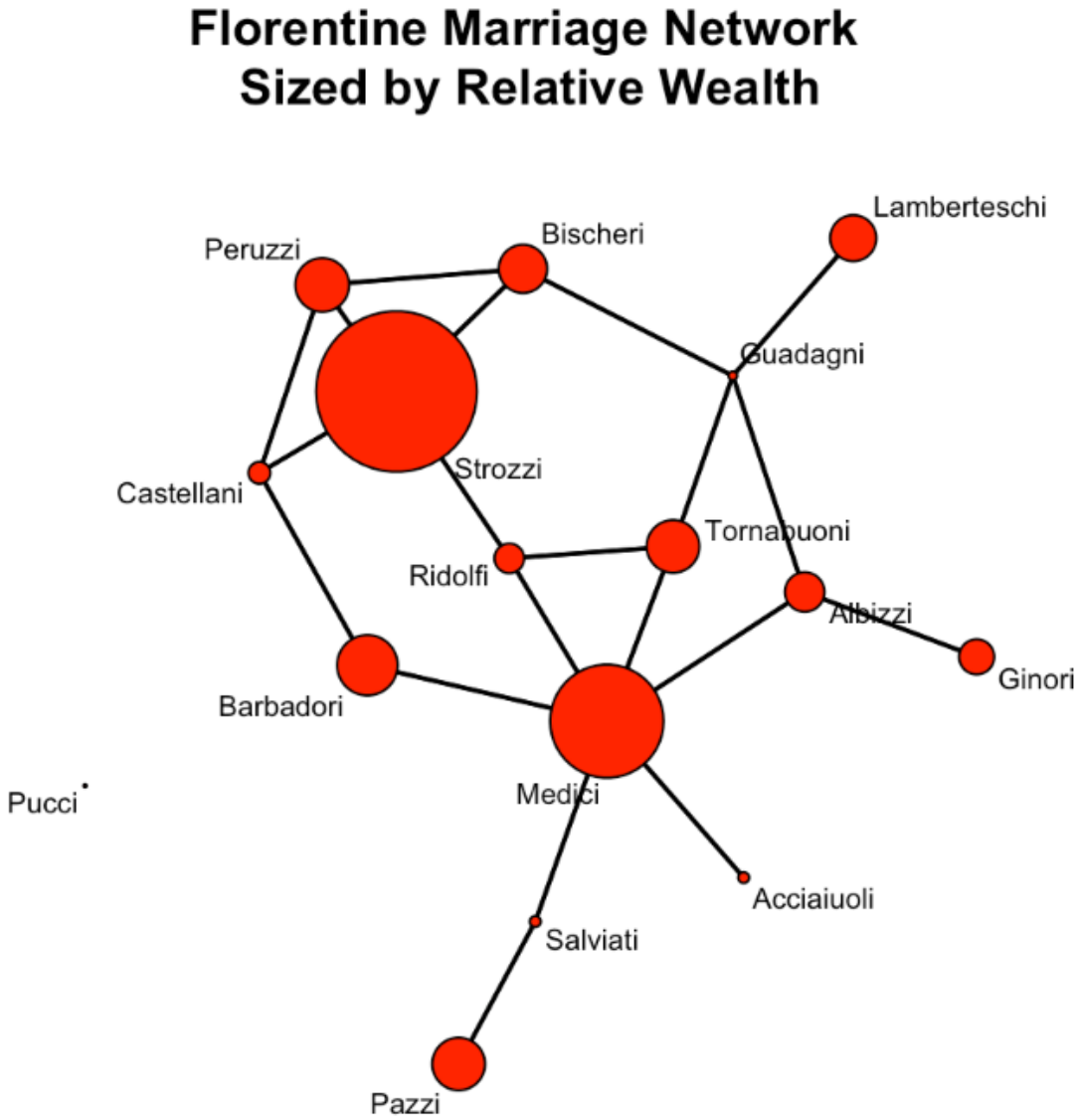
Analiza de rețea

- Cartografiază rețele pe baza **conexiunilor** dintre noduri
- Identifică proprietățile emergente ale rețelelor
  - De ex.: autocorelația sau homofilia de rețea; densitatea
    - „Cine se-aseamănă se-adună”/ „Birds of a feather fly together”
    - Filter bubble
- Identifică proprietățile nodurilor în rețea
  - De ex. centralitatea
- Identifică clustere de noduri bazate pe interconectivitate

Autocorelația

În ce măsură suntem influențați de cei similari cu noi – sau de noi înșine în trecut?

- Cine se-aseamănă se-adună
- Așchia nu sare departe de trunchi
- Orașele mai bogate se învecinează cu orașe mai bogate sau mai sărace?
- Familiile mai bogate se căsătoresc cu familii mai bogate sau mai sărace?
- Tendințe inerțiale – serii de timp

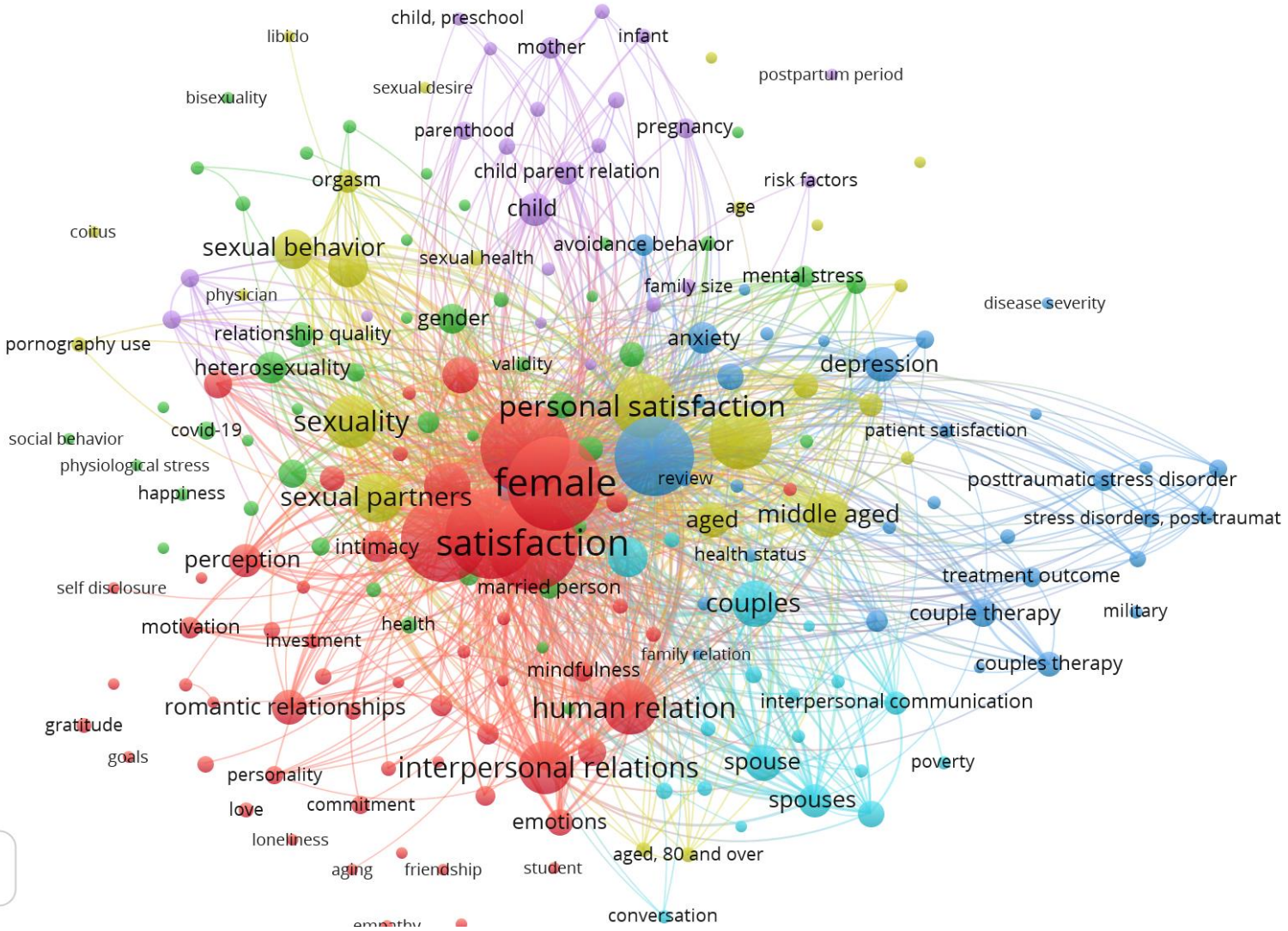


Sursa

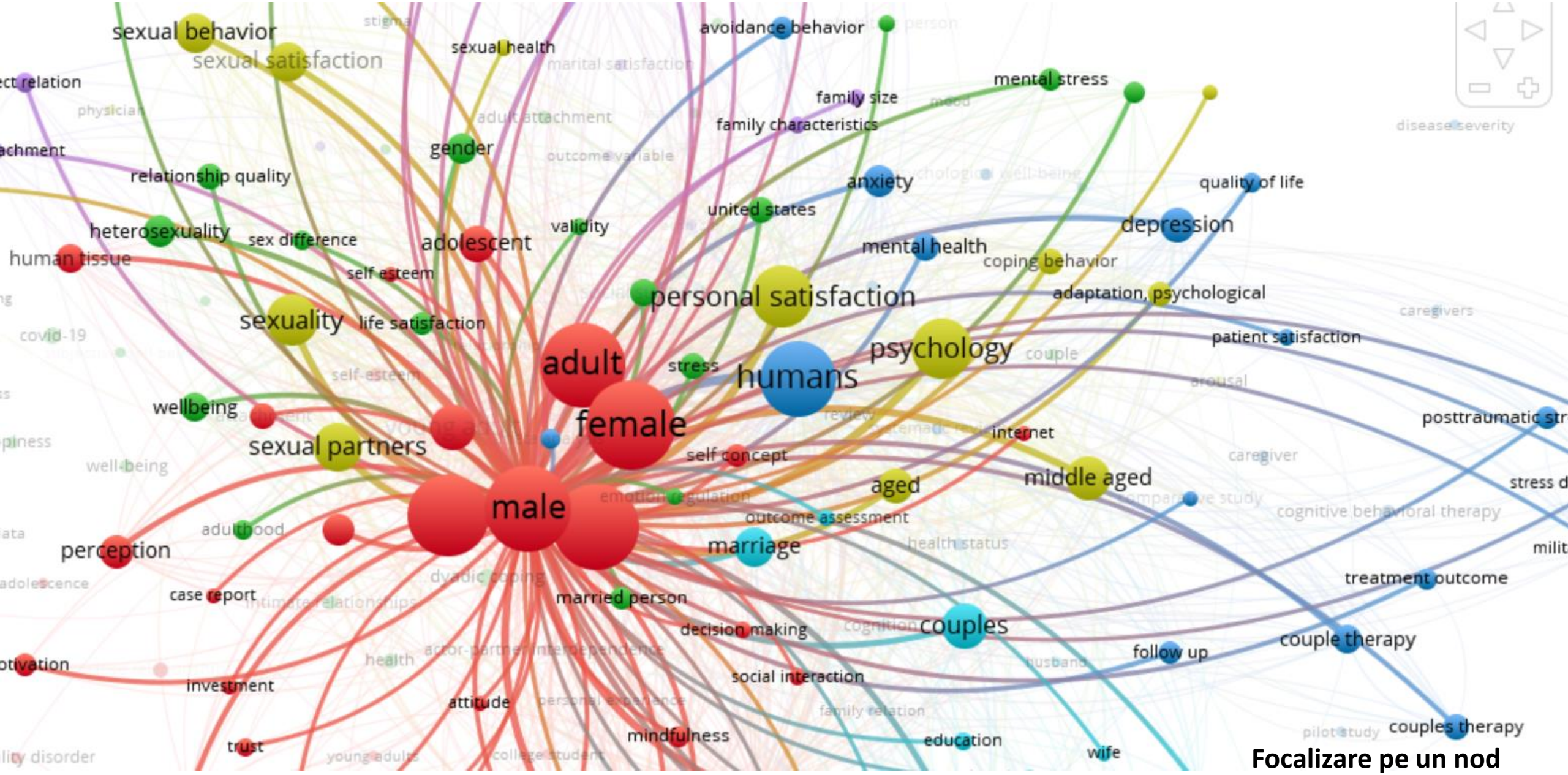
Co-ocurența

VOSviewer: hărți bibliometrice de co-ocurență a cuvintelor cheie

- Relationship satisfaction
- Scopus
- Ultimii 5 ani
- Social sciences și Psychology



Vizualizare a rețelei



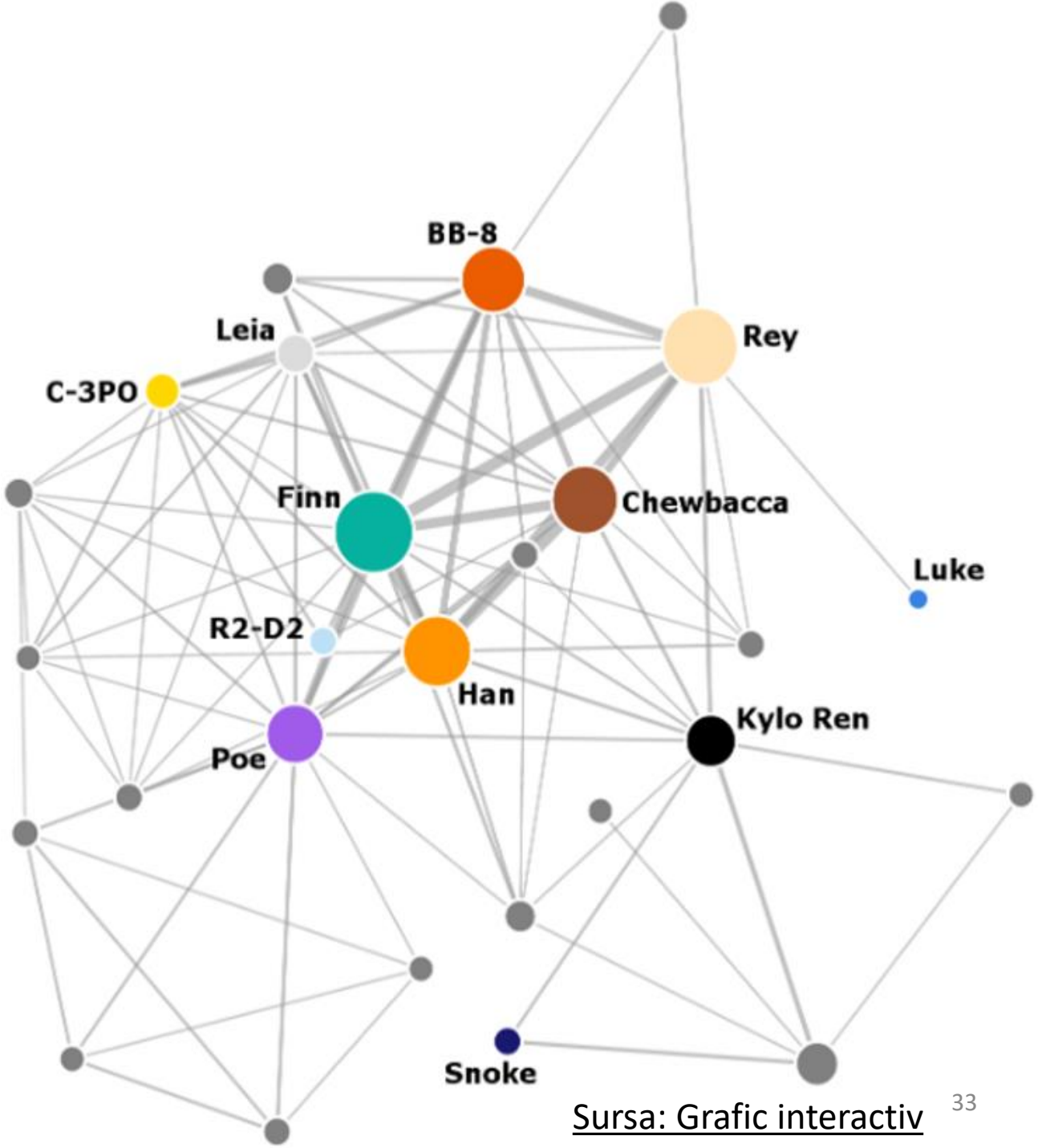
Focalizare pe un nod

Centralitatea

Cât de central / periferic este un nod în rețea?

Măsuri ale influenței / relevanței

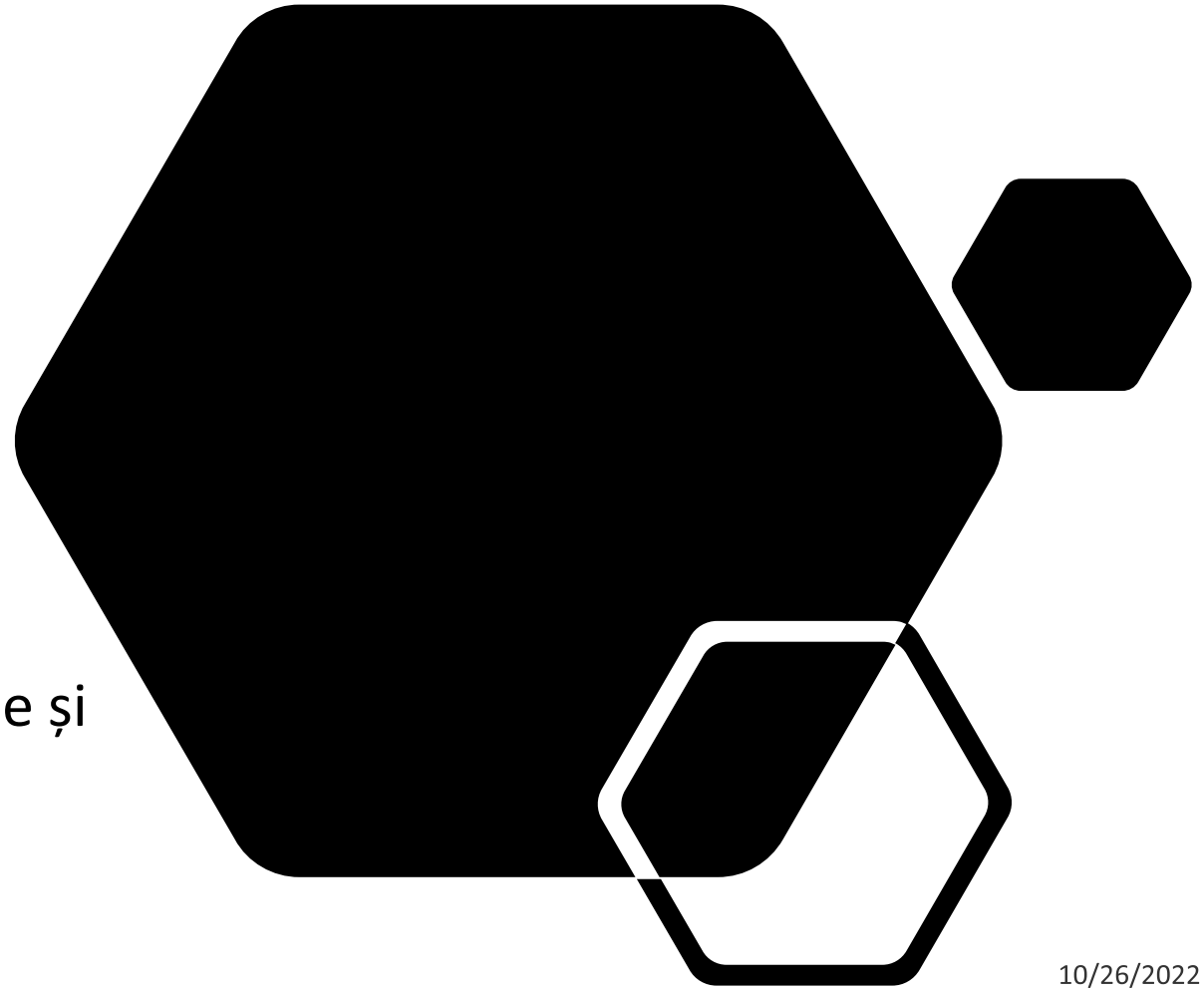
Ex: Analiza de rețea a personajelor din Star Wars ep VII



Sursa: Grafic interactiv

Serii de timp

Extrapolări și explicații prin tendințe temporale și factori externi



10/26/2022

Serii de timp

- Modelează variația unui fenomen ca funcție temporală
- Izolează și estimează tipuri de variații în timp
  - Tendințe, ciclicități, variații aleatorii
  - Alți factori externi influenți cu care fenomenul co-variază

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 intrinseci: cicluri, tendințe
  - Modele explicative extrinseci: cauze externe

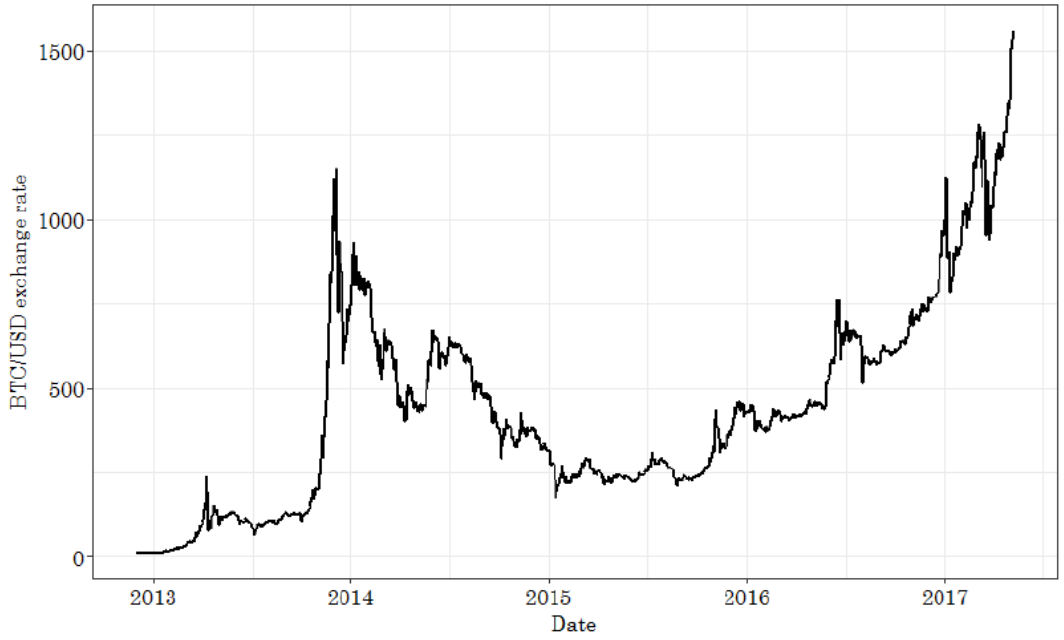


Figure 2: Bitcoin exchange rate with USD

Poyser 2018

Descompunerea în componente

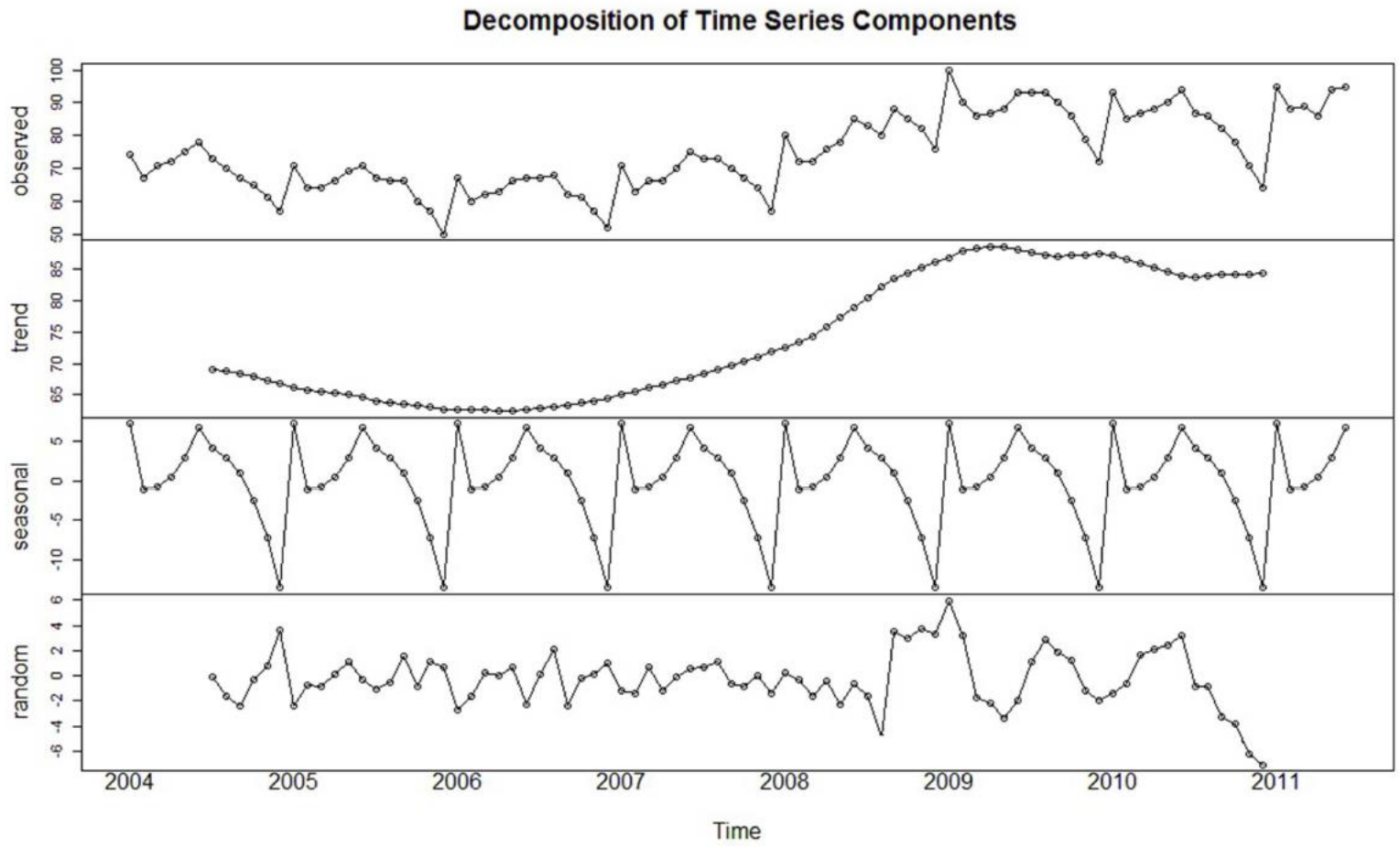
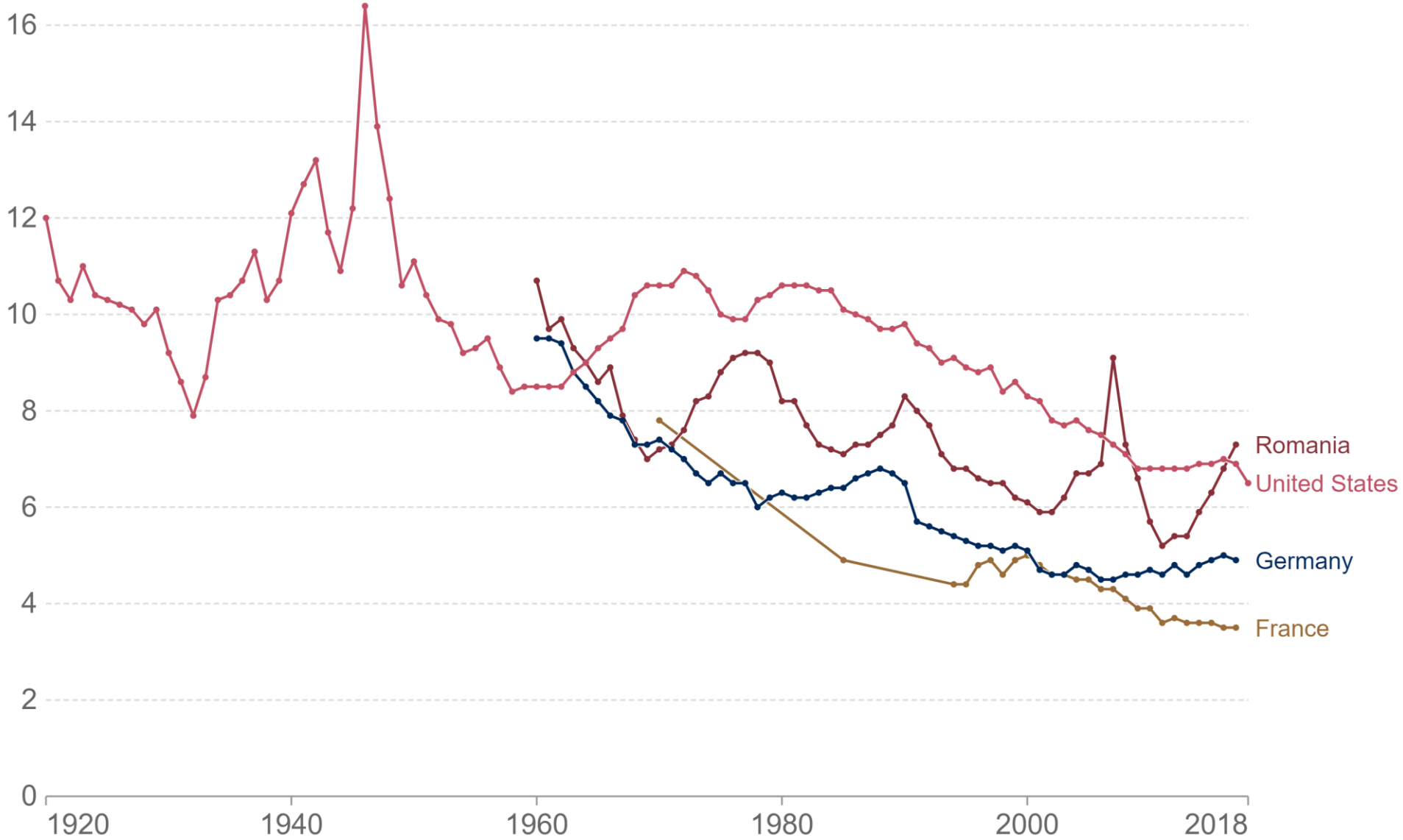


FIGURE 2 | The original time series decomposed into its trend, seasonal, and irregular (i.e., random) components. Cyclical effects are not present within this series. Sursa

Marriages per 1,000 people

Number of marriages in each year per 1,000 people in the population

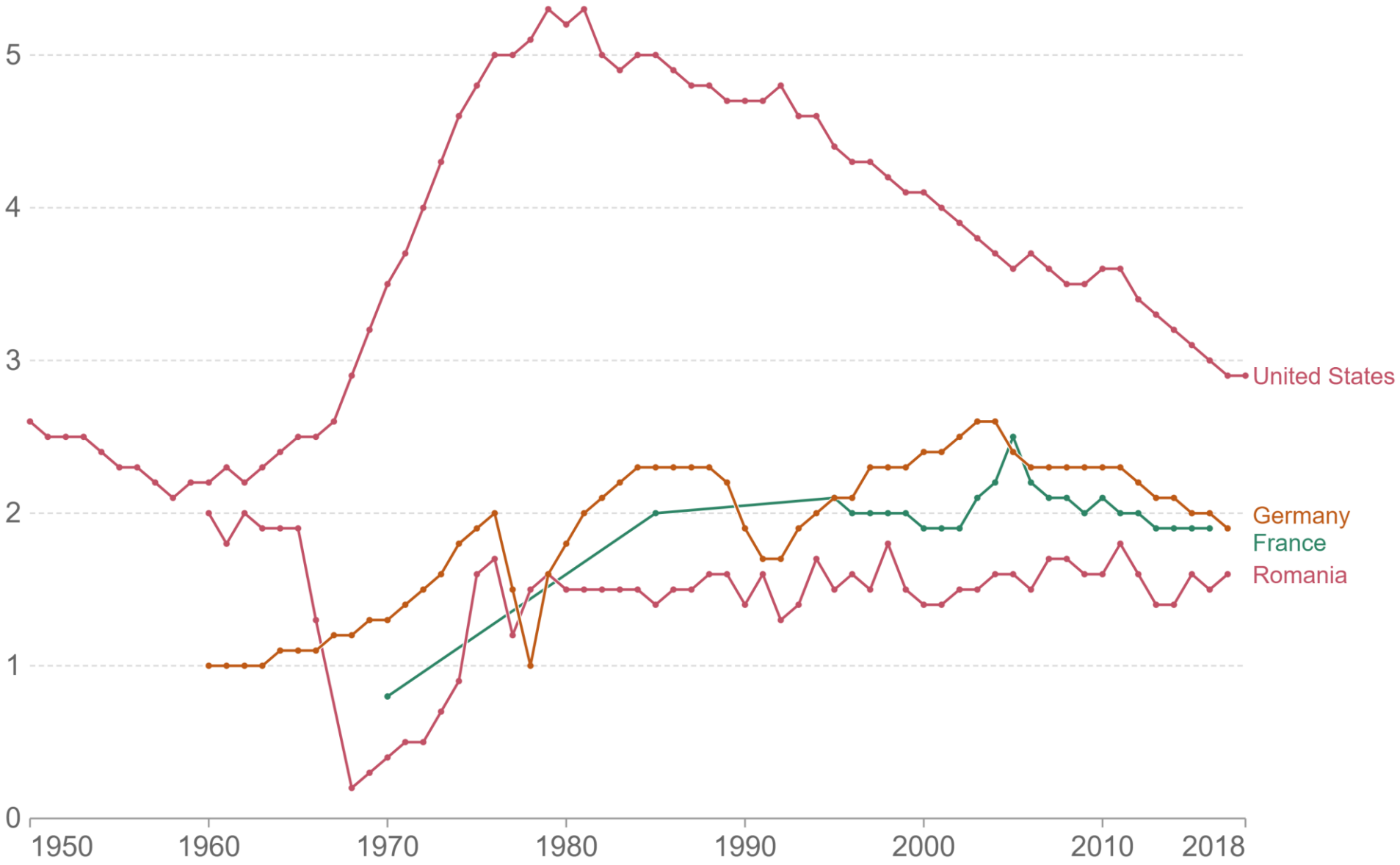


Source: OWID based on UN, OECD, Eurostat and others

OurWorldInData.org/marriages-and-divorces • CC BY

Sursa

Divorces per 1,000 people

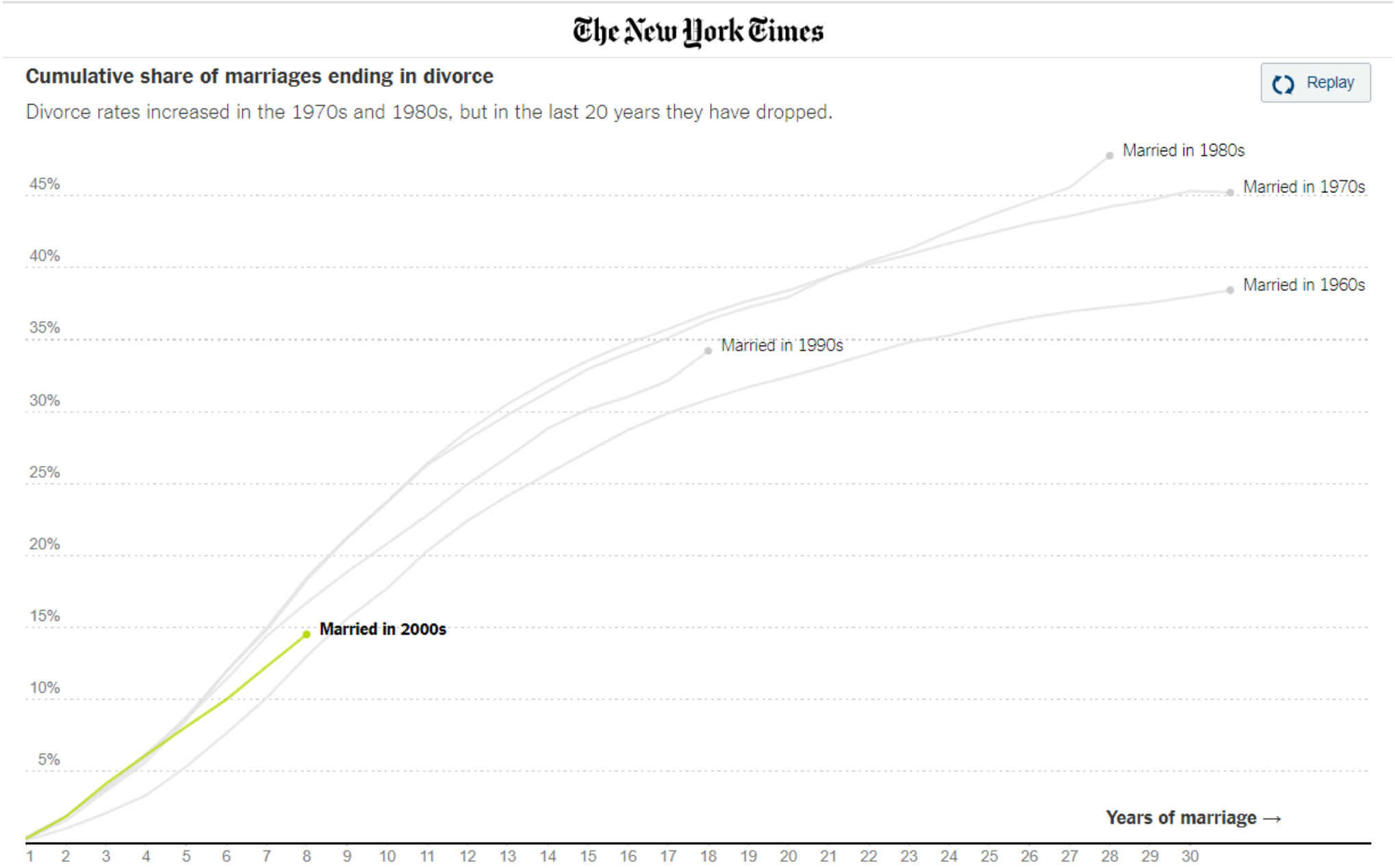


Source: OWID based on UN, OECD, Eurostat and other sources

OurWorldInData.org/marriages-and-divorces • CC BY

Sursa

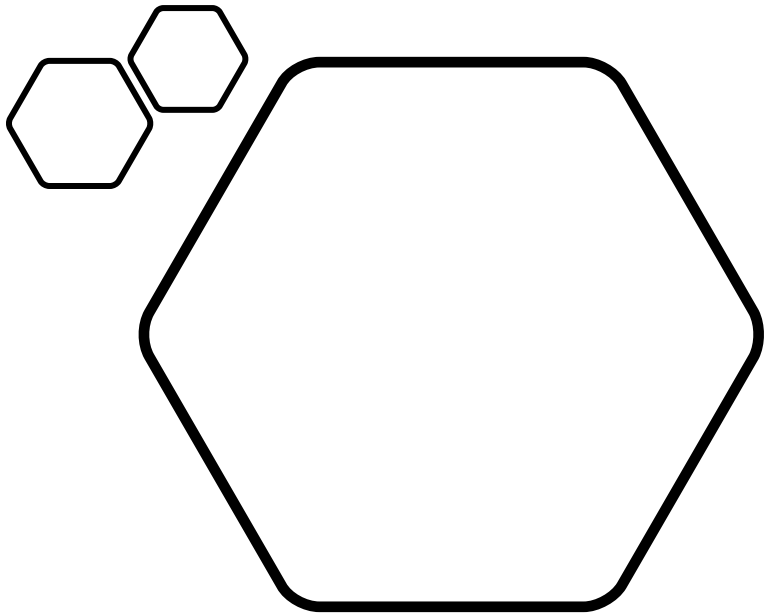
Ratele cumulative sunt mai stabile



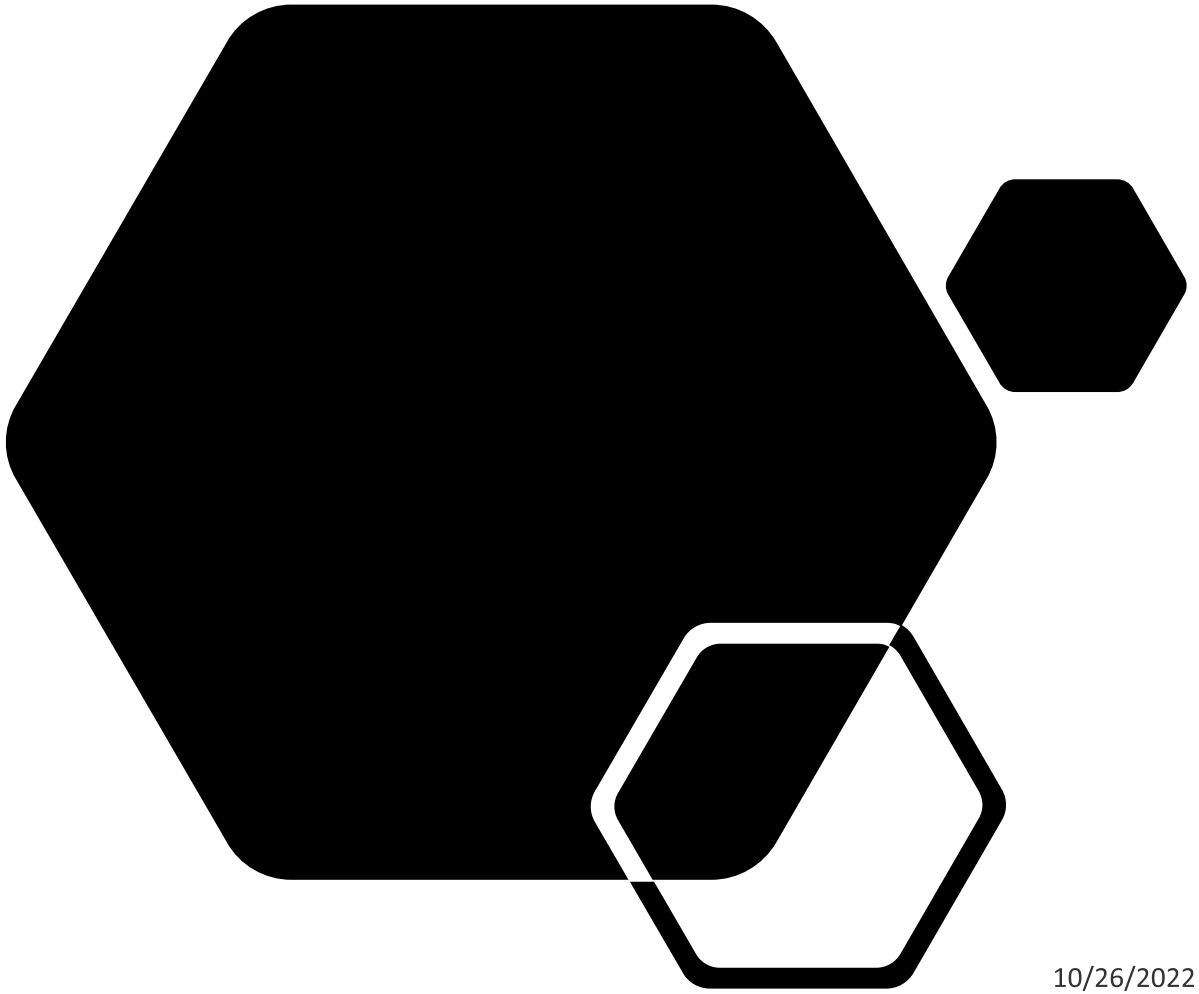
Sursa

Concluzii

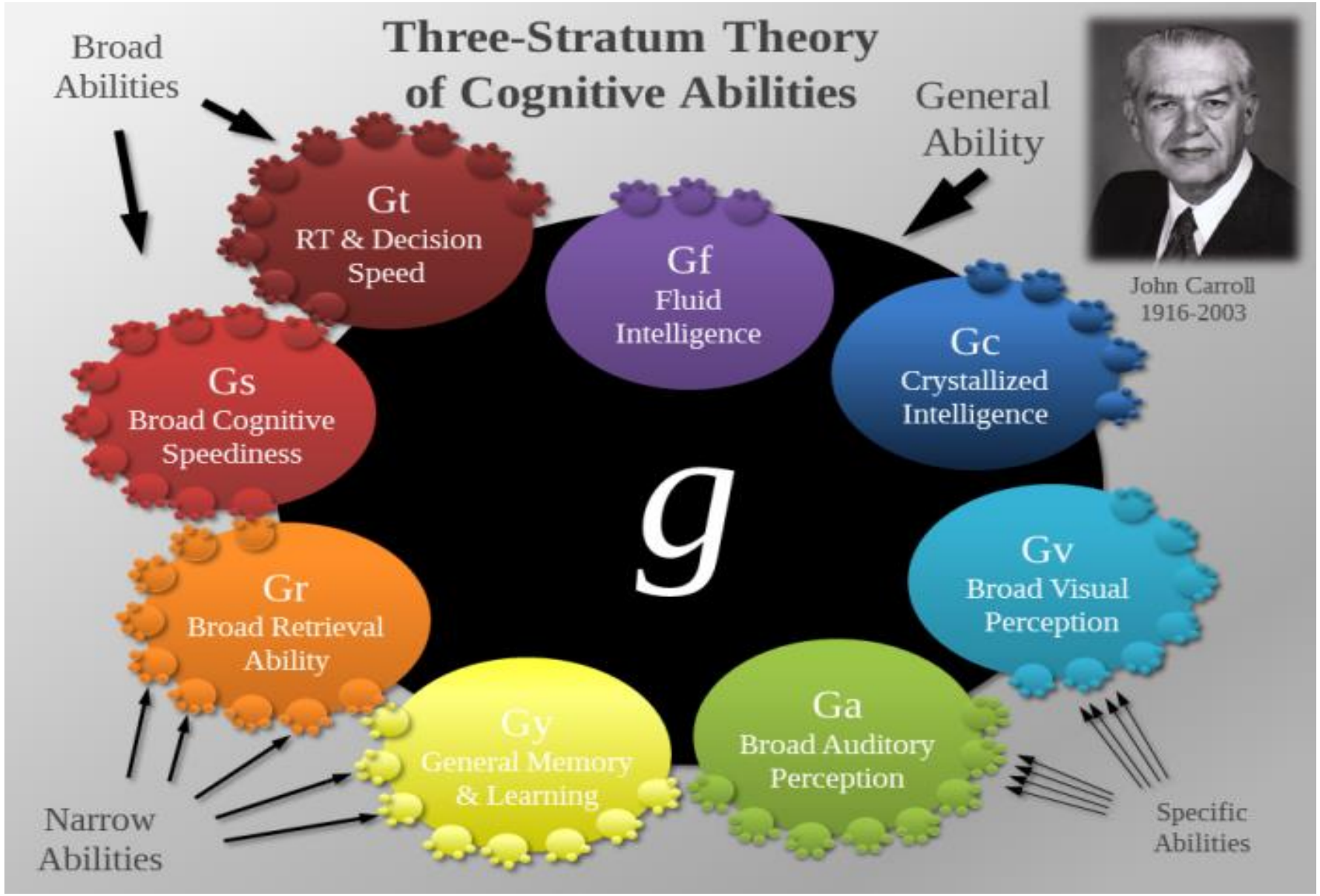
- Tehnicile de analiză reduc complexitatea datelor empirice
- Identificarea patternurilor relevante
- Scopuri
  - Explorarea: frecvențe, medii, corelații
  - Măsurare: reducerea dimensionalității
  - Clasificarea: analiza cluster
  - Explicarea: analiza de regresie, analiza de rețea
  - Extrapolarea: serii de timp



Extra time



Inteligențe multiple



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

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

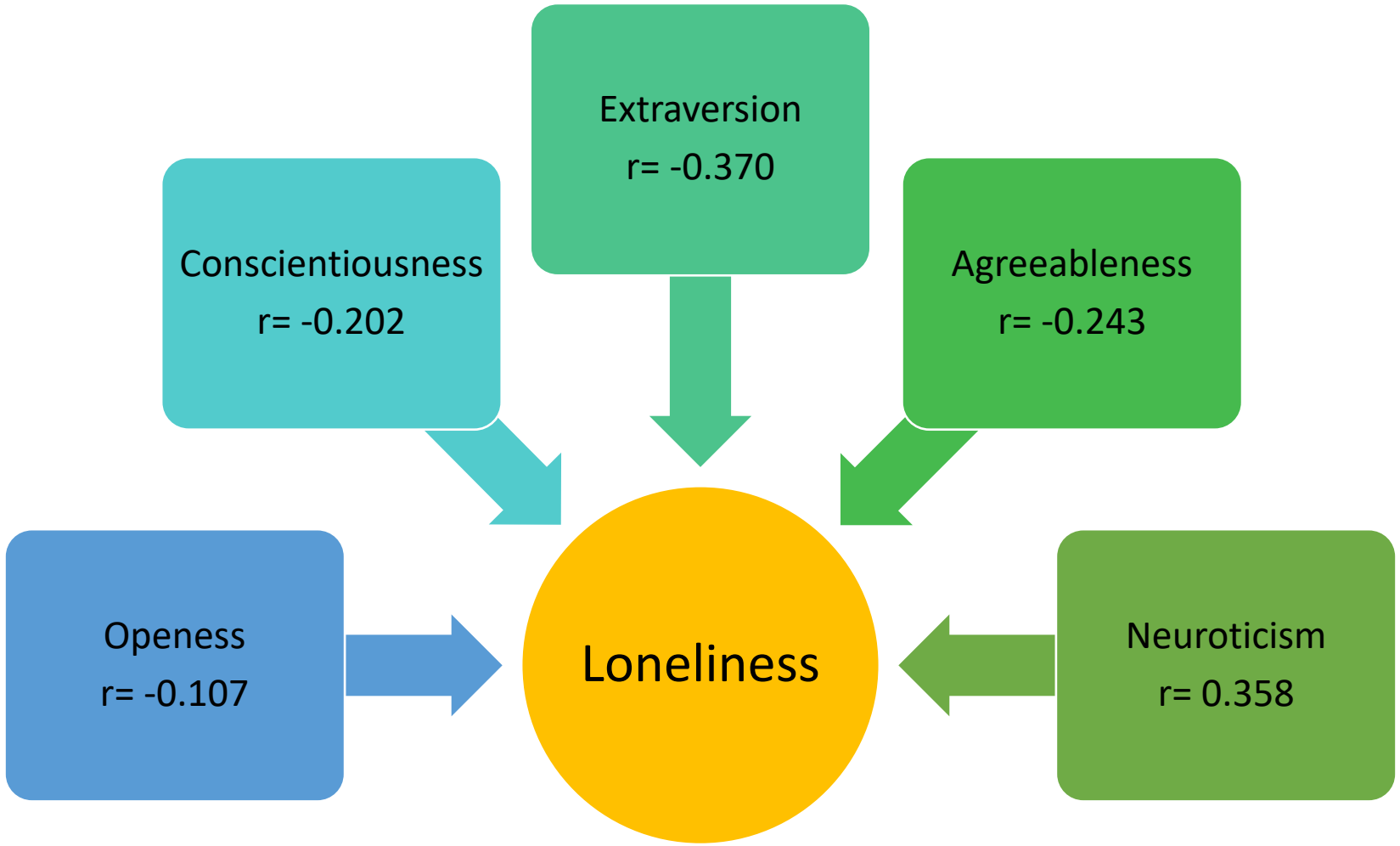
SUSANNE BUECKER[1]*, MARLIES MAES[2], JAAP J. A. DENISSEN[3] and MAIKE LUHMANN[1]

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

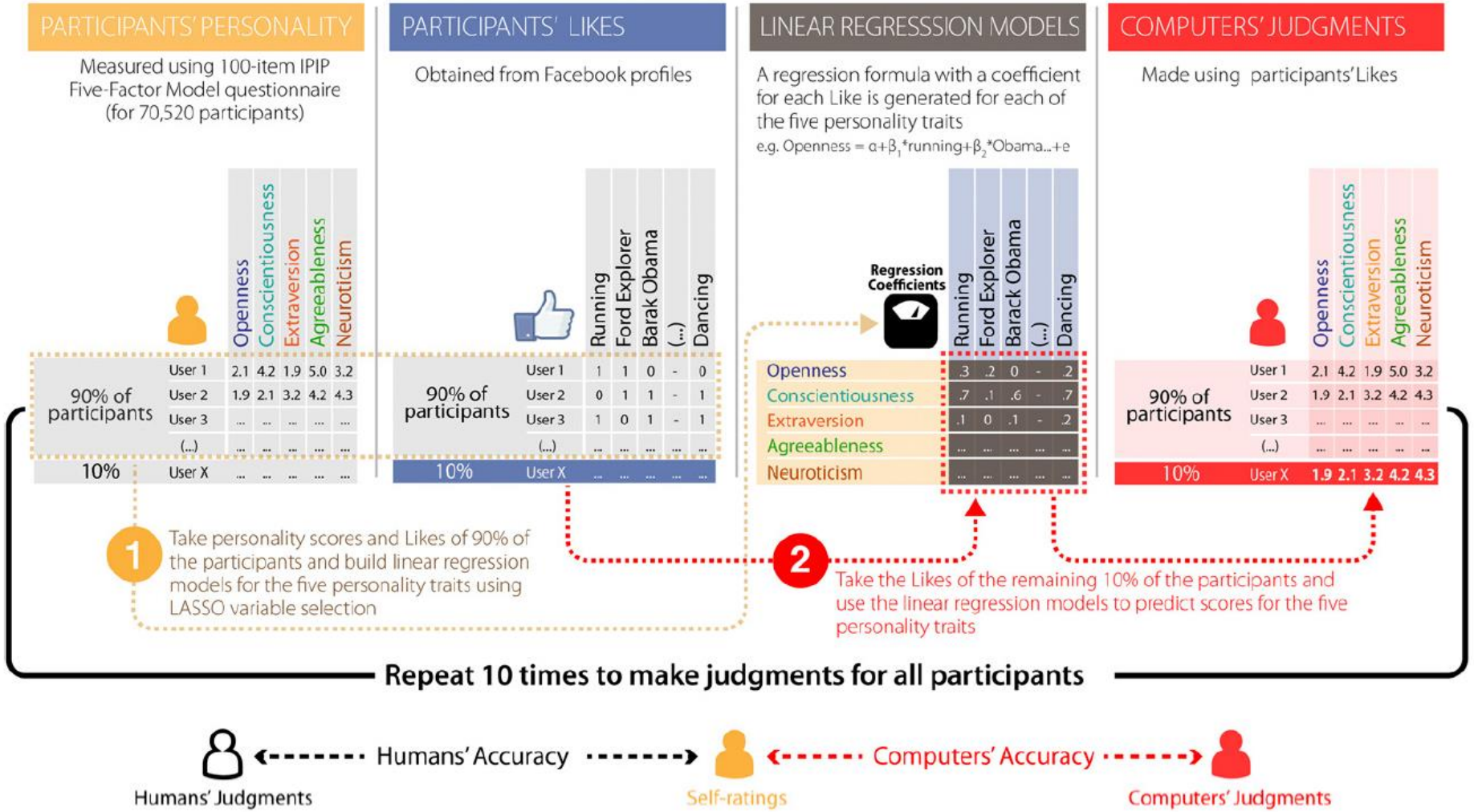
Abstract: This preregistered meta-analysis (k = 113, total n = 93 668) addressed how the Big Five dimensions of personality (extraversion, agreeableness, conscientiousness, neuroticism, and openness) are related to loneliness. Robust variance estimation accounting for the dependency of effect sizes was 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.

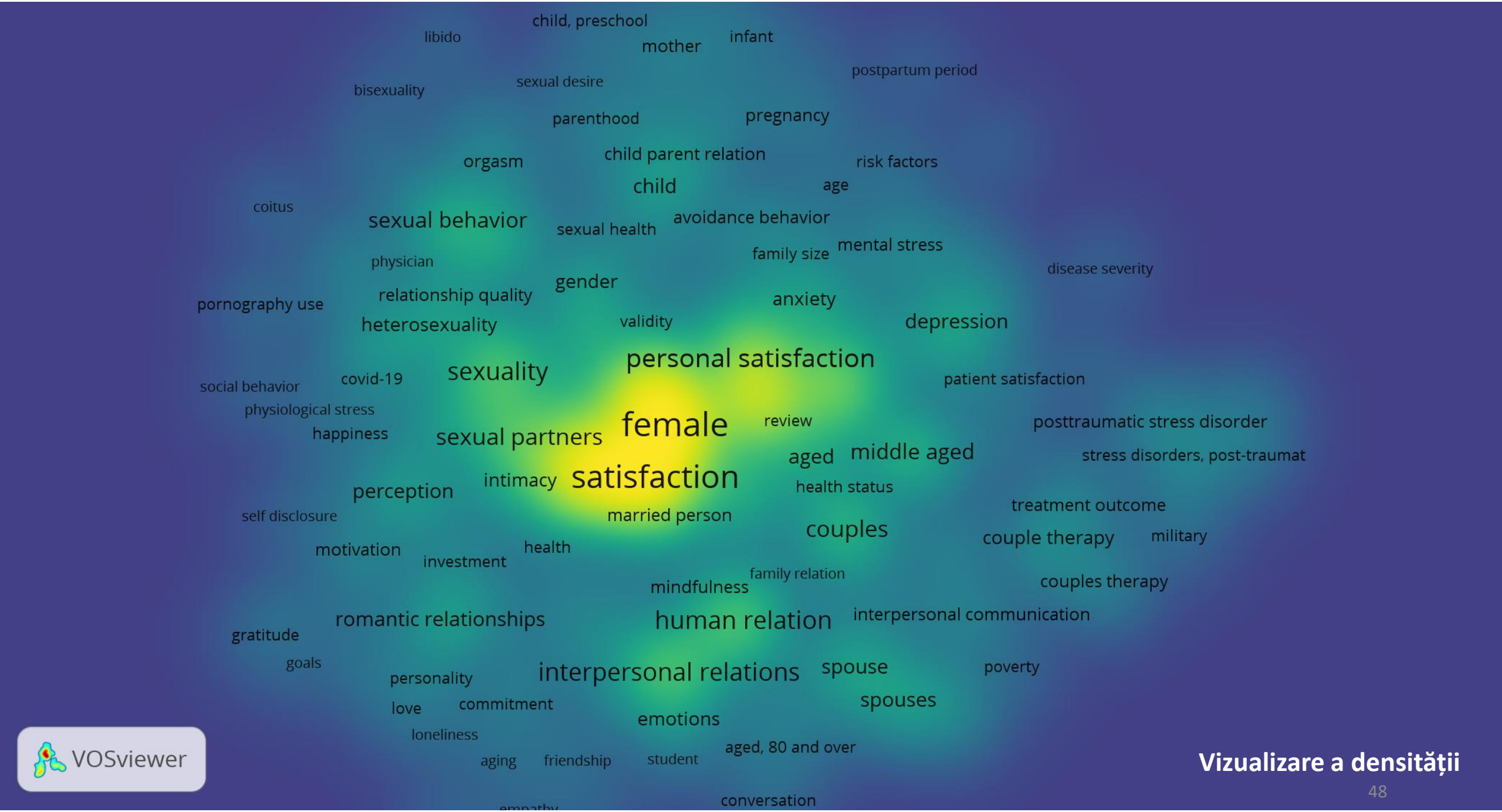
Key words: loneliness; perceived social isolation; personality; Big Five; meta-analysis

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

OCEAN și digital prediction



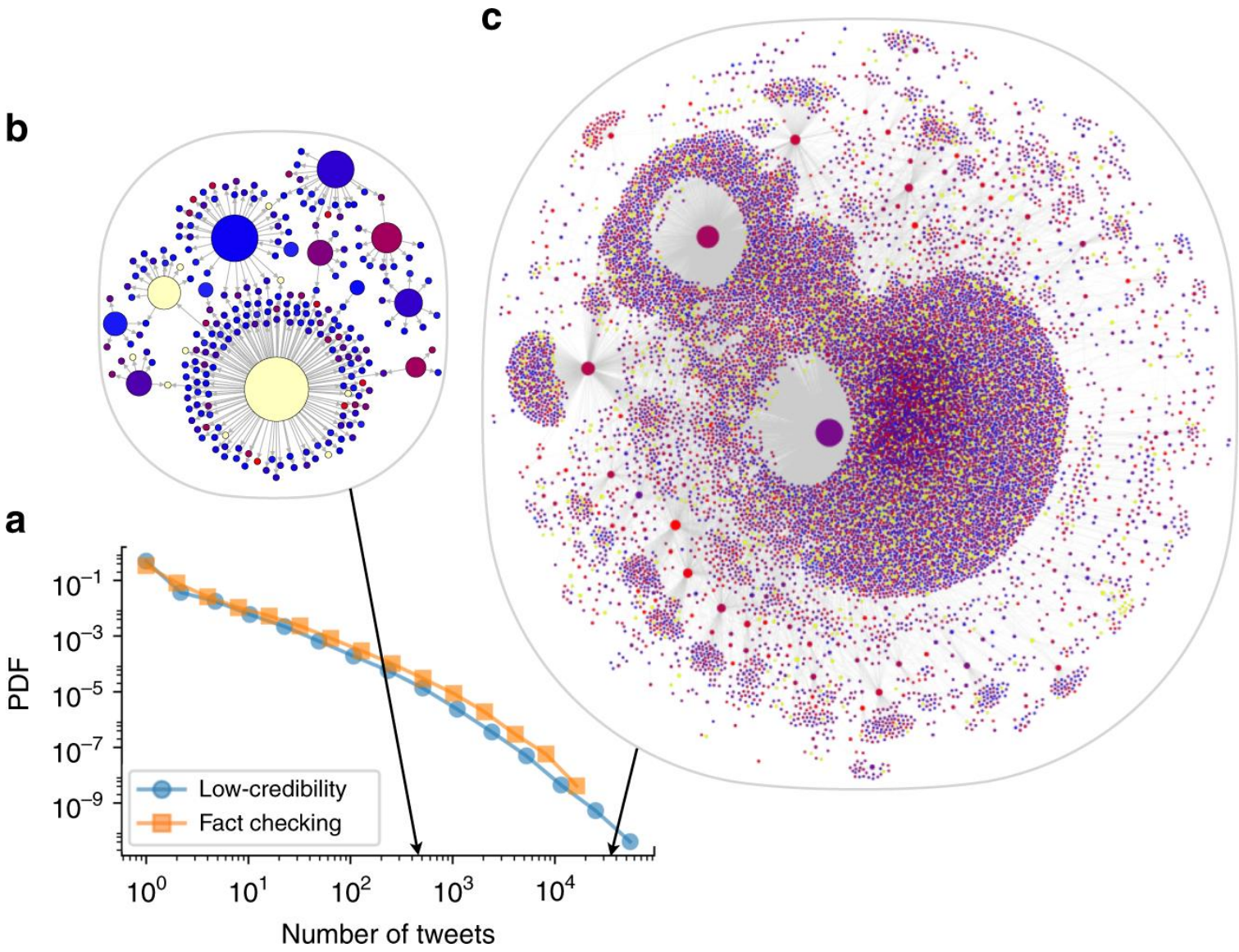


VOSviewer

Vizualizare a densității

Bots in fake news retweet networks

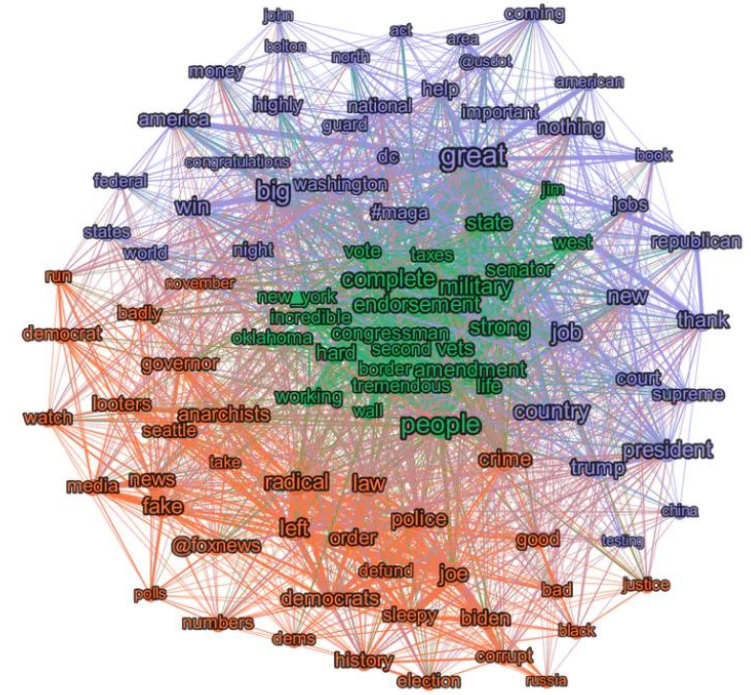
- As illustrations, the diffusion networks of two stories are shown: **b** a medium-virality misleading article titled “FBI just released the Anthony Weiner warrant, and it proves they stole election”, published a month after the 2016 US election and **shared in over 400 tweets**; and **c** a highly 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

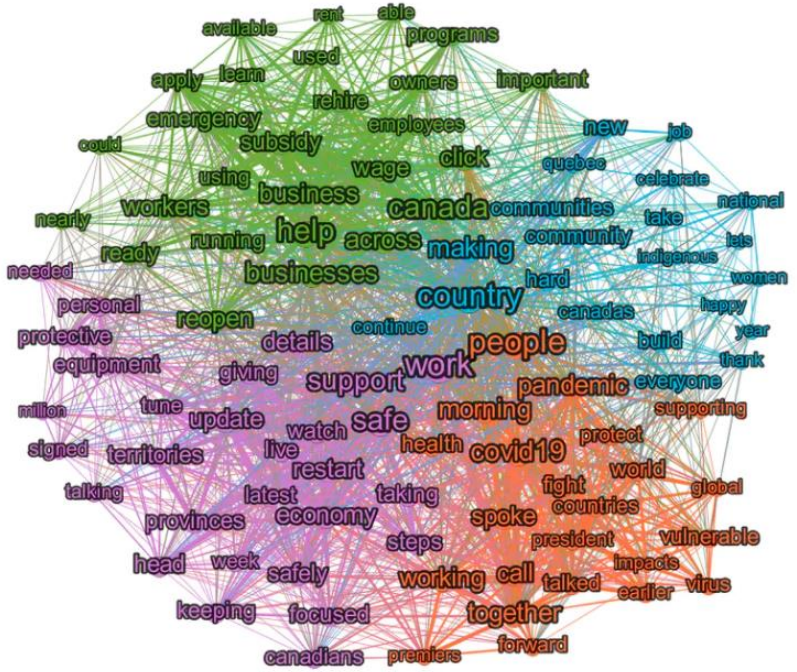
Rețele de co-ocurență pe Twitter

Trump, June 2020



Trump's June Twitter keyword network.

Trudeau, June 2020



Trudeau's June Twitter keyword network.

Bonato & Nazareth, 2020

The structure of videogame preference

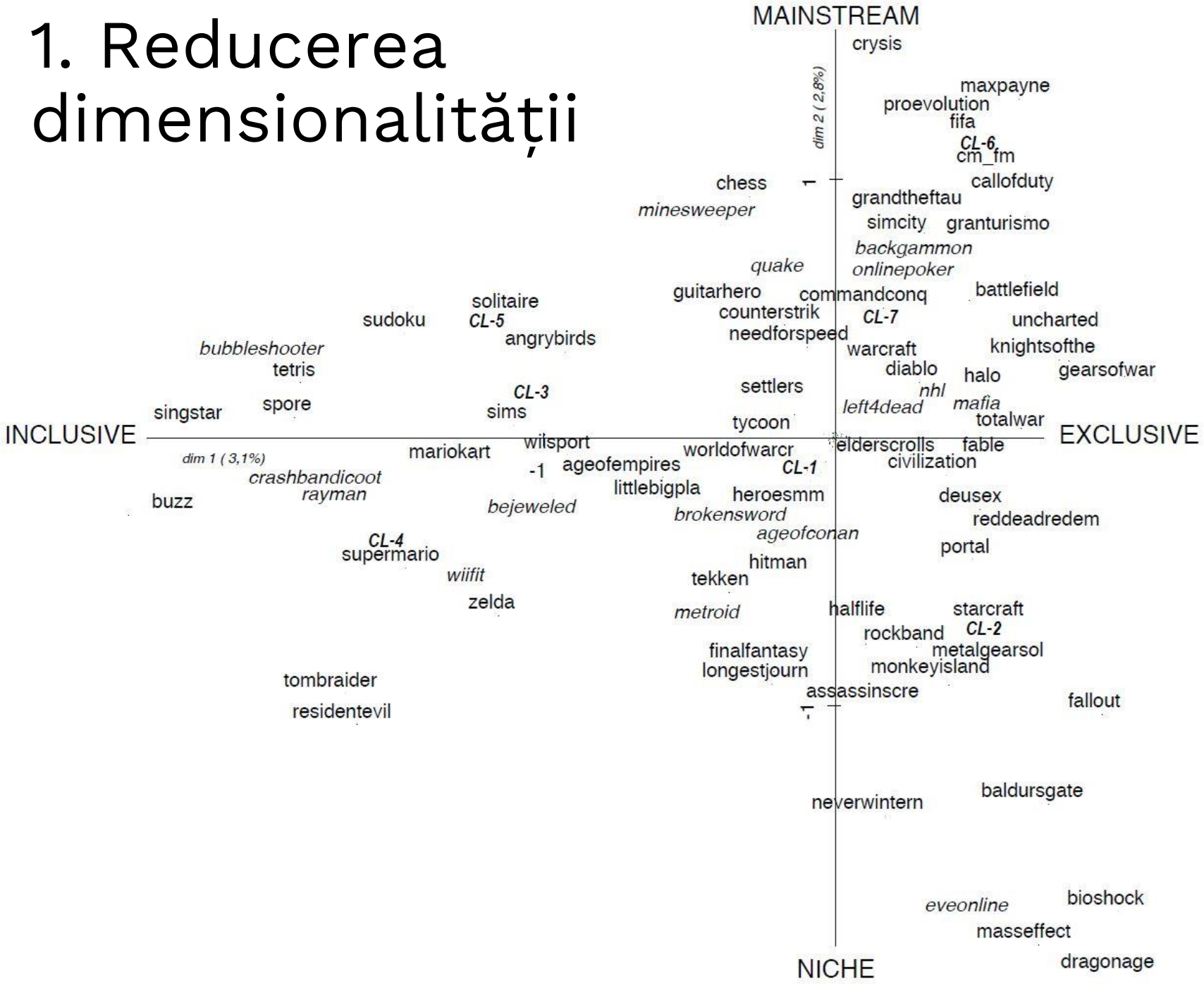
- Data: two broad surveys of Norwegian higher education students' cultural- media- and lifestyle preferences (2008, 2010)
- Of 2367 respondents, 758 students mentioned one or more favourite games, and 533 different titles were mentioned in total
  - Grouped in 328 game series, of which 64 chosen by 5+ 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%

	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 %**