HTE, 2022. október 12.

Mindennek mértéke az ember.
(Prótagorasz)

Az emberközpontú mesterséges intelligencia műszaki kérdései

Antal Péter
Bioinformatikai Laboratórium
Mesterséges Intelligencia Csoport
Méréstechnika és Információ Rendszerek Tanszék
Villamosmérnöki és Informatikai Kar
Budapesti Műszaki és Gazdaságtudományi Egyetem

Ágenda

- Az emberközpontú mesterséges intelligencia (EMMI)
- MI dióhéjban és a standard MI model
- MI trendek, általános MI, szuperMI
- Az MI veszélyei
- EMMI megoldások
  - Bizonyíthatóan jóravaló MI
  - Értelmezhető MI
  - Magyarázható MI
  - Federált MI

Emberközpontú mesterséges intelligencia

HCAIM-projekt: Mesterfokon, etikusan az MI-ről

- 60 kredites EU szintű HCAIM program
- Partnerek:
- TU Dublin, HU Utrecht, Federico II Napoli, BME
  - + kutatóintézetek
  - + KKV-k
- Projekthonlapok:
  - közös: http://humancentered-ai.com/
  - BME: https://hcaim.bme.hu

human centred artificial intelligence masters Budapest

https://computerworld.hu/technologia/hcaim-projekt-mesterfokon-etikusan-a-mesterseges-intelligenciarol-310148.html

Emberközpontú MI definiciója (és szabályozások)

- AI HLEG
  - "The human-centric approach to AI strives to ensure that human values are central to how AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights."
  - **A mesterséges intelligencia emberközpontú megközelítése arra törekszik, hogy az alapvető jogok tiszteletben tartása révén a humán értékek központi szerepet kapjanak az MI-rendszerek fejlesztése, telepítése, használata és felügyelete során.**
- AI regulations.
  - General Data Protection Regulation (GDPR): Regulation (EU) 2016/679
  - Building Trust in Human Centric Artificial Intelligence: COM(2019)168
  - AI regulation: COM/2021/206 final1

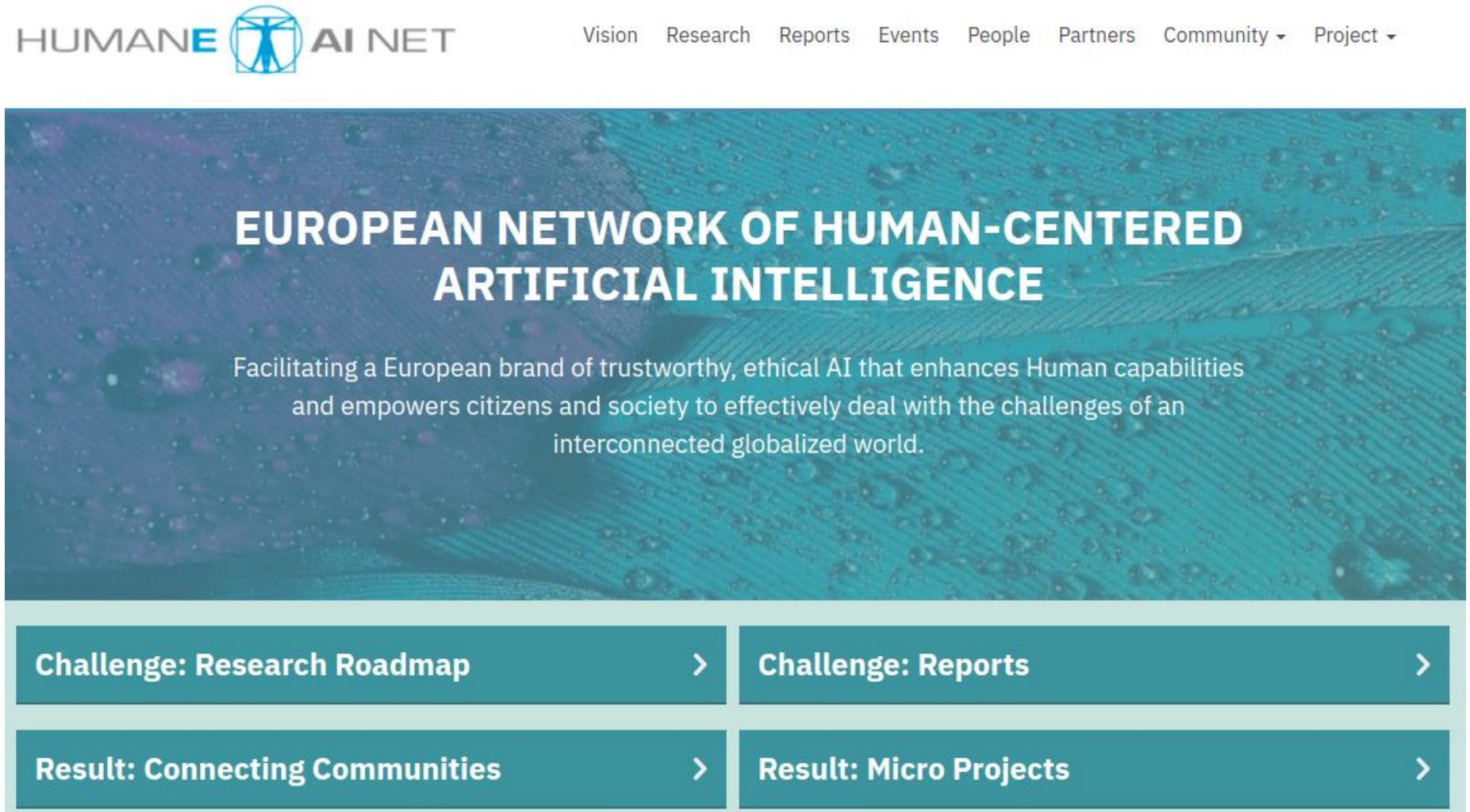
A HCAI értelmezés a HCAIM-ban

- **Beneficial AI:** AI solutions for humans, societies, and mankind
- **Trustworthy AI:** value compatibitility, understandability
- **AI safety**: formal methods in systems engineering.
- **Personal and institutional autonomy and freedom**: data security and privacy.
- **HC data analysis by design**: MLOps, prior knowledge, explanation, active learning, machine teaching in cooperative intelligence, automated data analytics.
- **HC knowledge engineering**: coding systems, ontologies, linked open data, summary statistics, automation of science (life sciences).
- **Improved cognitive enhancers:** personal assistants in education, intelligent citizen and costumer services, and decision support tools in personalized medicine.
- **Improved sensorial man-machine interfaces:** improved communication (speech, augmented reality) and human-computer interaction.
- **Improved sensorimotoric man-machine cooperation**: robotics, health-care assistance using wearable electronic devices, and in automated driver assistance systems/autonomous vehicles.
- **Smart devices, smart cities**: IoT, sensor fusion for predictive maintenance.
- ~~**Autonomous vehicles**~~

CLASSIC ("existential risk")

HCAIM

HCI: human-computer interaction



https://www.humane-ai.eu/

Források emberközpontú MI-hez

- AI25@ BME VIK
  - **Stuart Russel: A New Approach to AI, Budapest, 2018**
    - https://ai25.mit.bme.hu/en/program/russell/
    - https://ai25.mit.bme.hu/hu/program/
- **Stuart Russel: A Human-Centered Approach to Artificial Intelligence**
    - https://ucsdnews.ucsd.edu/feature/a-human-centered-approach-to-artificial-intelligence
- Mindscape
  - Derek Leben on Ethics for Robots and Artificial Intelligences
    - https://www.preposterousuniverse.com/podcast/2019/01/21/episode-30-derek-leben-on-ethics-for-robots-and-artificial-intelligences/
- Human-Centered AI: Keynotes and Panel by Yoshua Bengio and Ben Shneiderman
  - https://www.youtube.com/watch?v=uc0bYp_-JLA

MI dióhéjban és a standard MI model

What is AI?

Thinking humanly	Thinking rationally
Acting humanly	Acting rationally

Stuart Russel: "standard model"

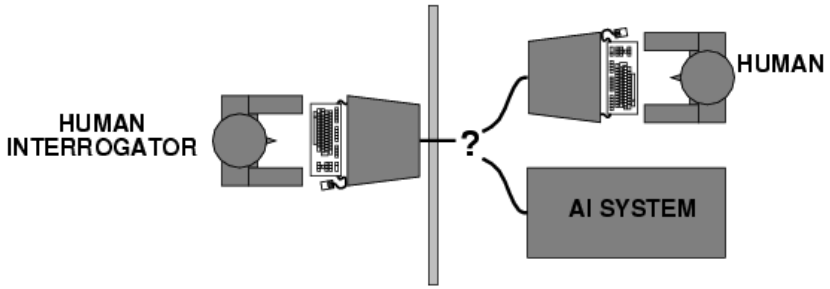
- **Machines are intelligent to the extent that their actions can be expected to achieve their objectives.**

Wang, Pei. "**On Defining Artificial Intelligence**." Journal of Artificial General Intelligence 10.2 (2019): 1-37.

- *The essence of intelligence is the principle of adapting to the environment while working with insufficient knowledge and resources. Accordingly, an intelligent system should rely on finite processing capacity, work in real time, open to unexpected tasks, and learn from experience. This working definition interprets "intelligence" as a form of "relative rationality" (Wang, 2008),*

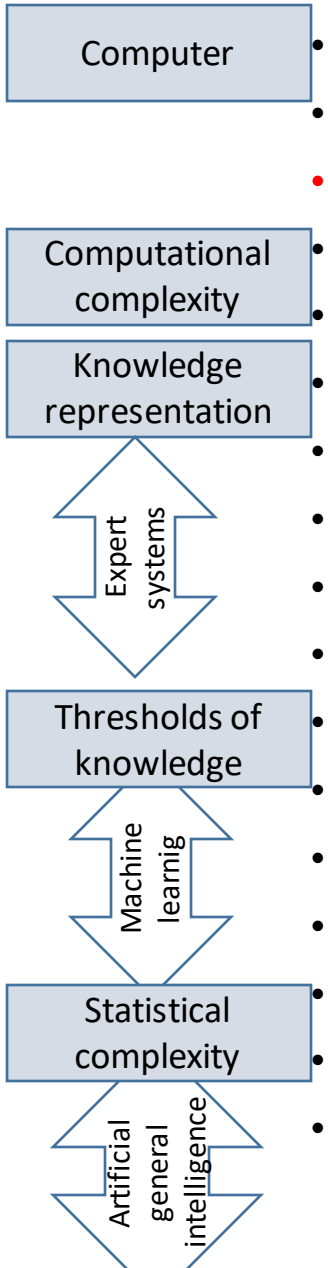
Acting humanly: Turing Test

- Turing (1950) "Computing machinery and intelligence":
- "Can machines think?" → "Can machines behave intelligently?"
- Operational test for intelligent behavior: the Imitation Game



- Predicted that by 2000, a machine might have a 30% chance of fooling a lay person for 5 minutes
- R. Kurzweil (1999): 2029
- Consensus AI expert opinion (2022): ~2030

Milestones and phases in AI/ML



- ~1930: Zuse, Neumann, Turing..: „instruction is data”:
- 1943 McCulloch & Pitts: Boolean circuit model of brain
- 1950 Turing's "Computing Machinery and Intelligence"
- 1956 Dartmouth meeting: the term "Artificial Intelligence”
- 1950s Early AI programs (e.g. Newell & Simon's Logic Theorist)
- **The physical symbol system hypothesis: search**
- 1965 Robinson's complete algorithm for logical reasoning
- 1966—73 AI discovers computational complexity
- 1969—79 Early development of knowledge-based systems
- **The knowledge system hypothesis: knowledge is power**
- 1986- Neural networks return to popularity
- 1988- Probabilistic expert systems/**causality research(2000/2020<)**
- 1990- Kernel methods
- **The „big data” hypothesis: let the data speak**
- 1995-2006/2009/2012 Emergence of machine learning
- 2010< Deep learning, GPUs,
- 2016<: reinforcement learning
- **2020<: human-centered/trustworthy/explainable/provably beneficial AI**

AIMA

A bayesi paradigma

Szubjektív valószínűségek

$$\lim_{N\to\infty}\frac{N_A}{N}=\lim_{N\to\infty}\hat{p}_N(A)=p(A)\,?\,p(A\,|\,\xi)$$

Bayes szabály

$$p(X\,|\,Y)=\frac{p(Y\,|\,X)\,p(X)}{p(Y)}=\frac{p(Y\,|\,X)\,p(X)}{\sum_X p(Y\,|\,X)\,p(X)}$$

$$p(Model\,|\,Data)\propto p(Data\,|\,Model)\,p(Model)$$

Thomas Bayes: 1701 – 1761

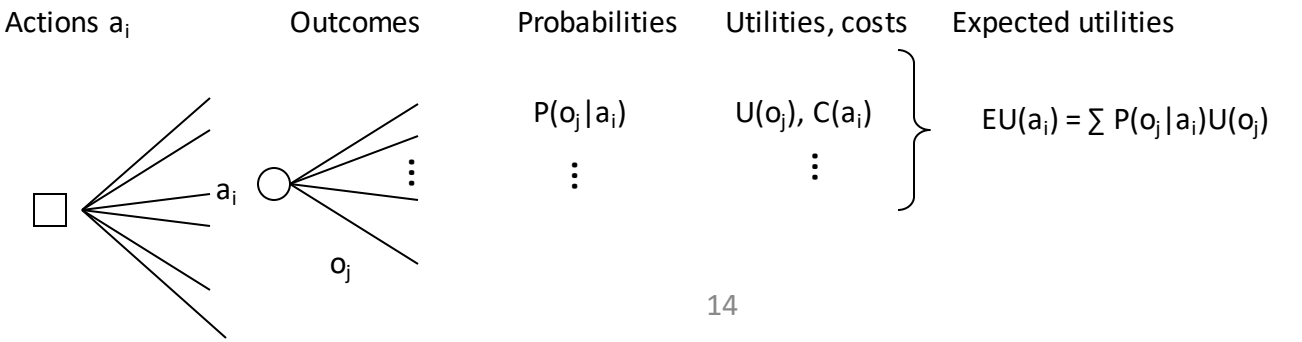
Bayesi modellátlagolás

$$p(X_{n+1}|X_{1:n})\propto\sum_m p(X_{n+1}|Model=m)p(X_{1:n}|Model=m)p(Model=m)$$

Solomonoff, Ray. "Complexity-based induction systems: comparisons and convergence theorems." *IEEE transactions on Information Theory* 24.4 (1978): 422-432.

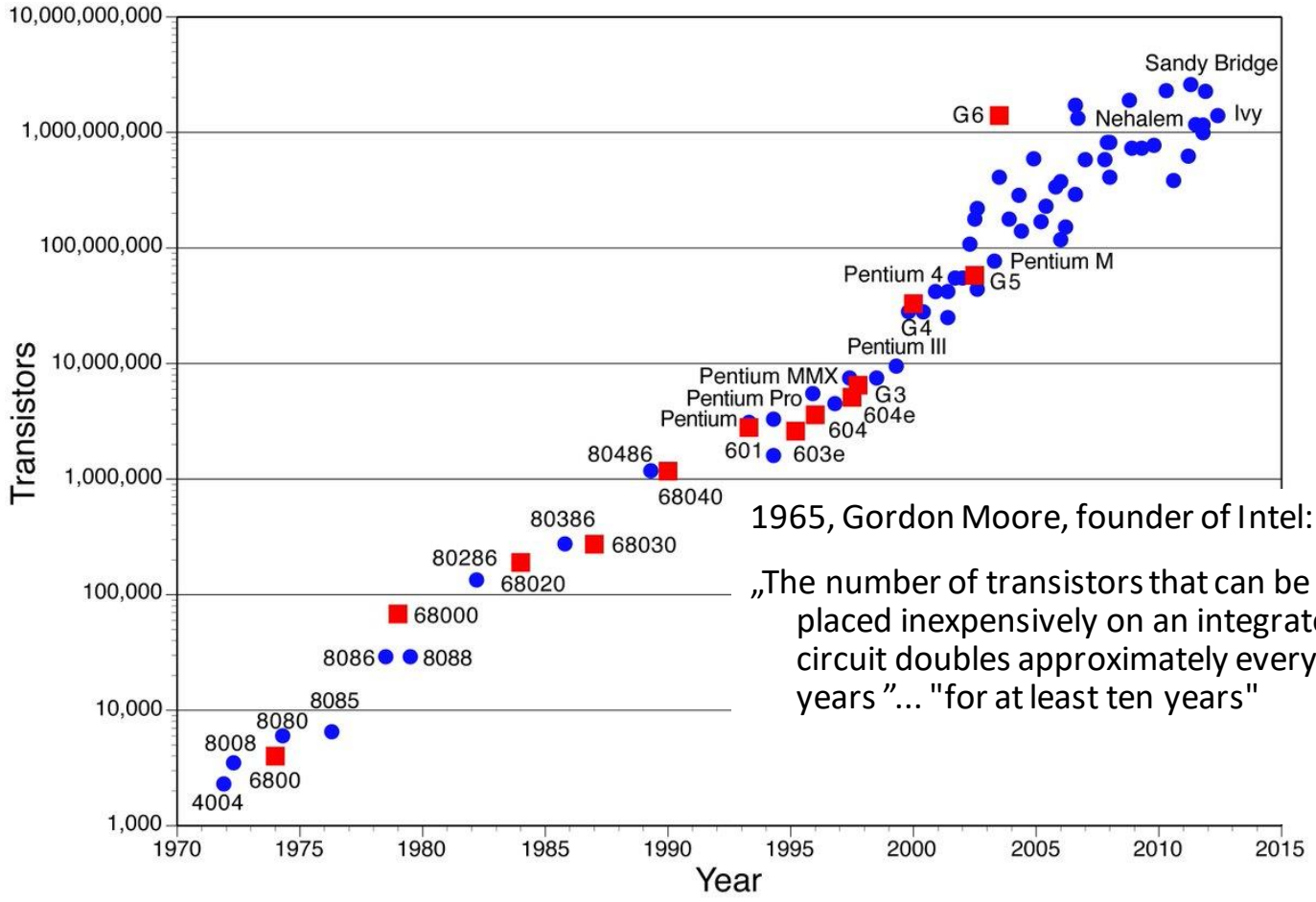
Bayesian decision theory probability theory+utility theory

- Decision situation:
  - Actions a_i
  - Outcomes o_j
  - Probabilities of outcomes $p(o_j \mid a_i)$
  - Utilities/losses of outcomes $U(o_j \mid a_i)$
  - Maximum Expected Utility Principle (MEU) $EU(a_i) = \sum_j U(o_j \mid a_i)\, p(o_j \mid a_i)$
  - Best action is the one with maximum expected utility $a^* = \arg\max_i EU(a_i)$



MI trendek, általános MI és a szuperMI

Computing power: Moore's Law



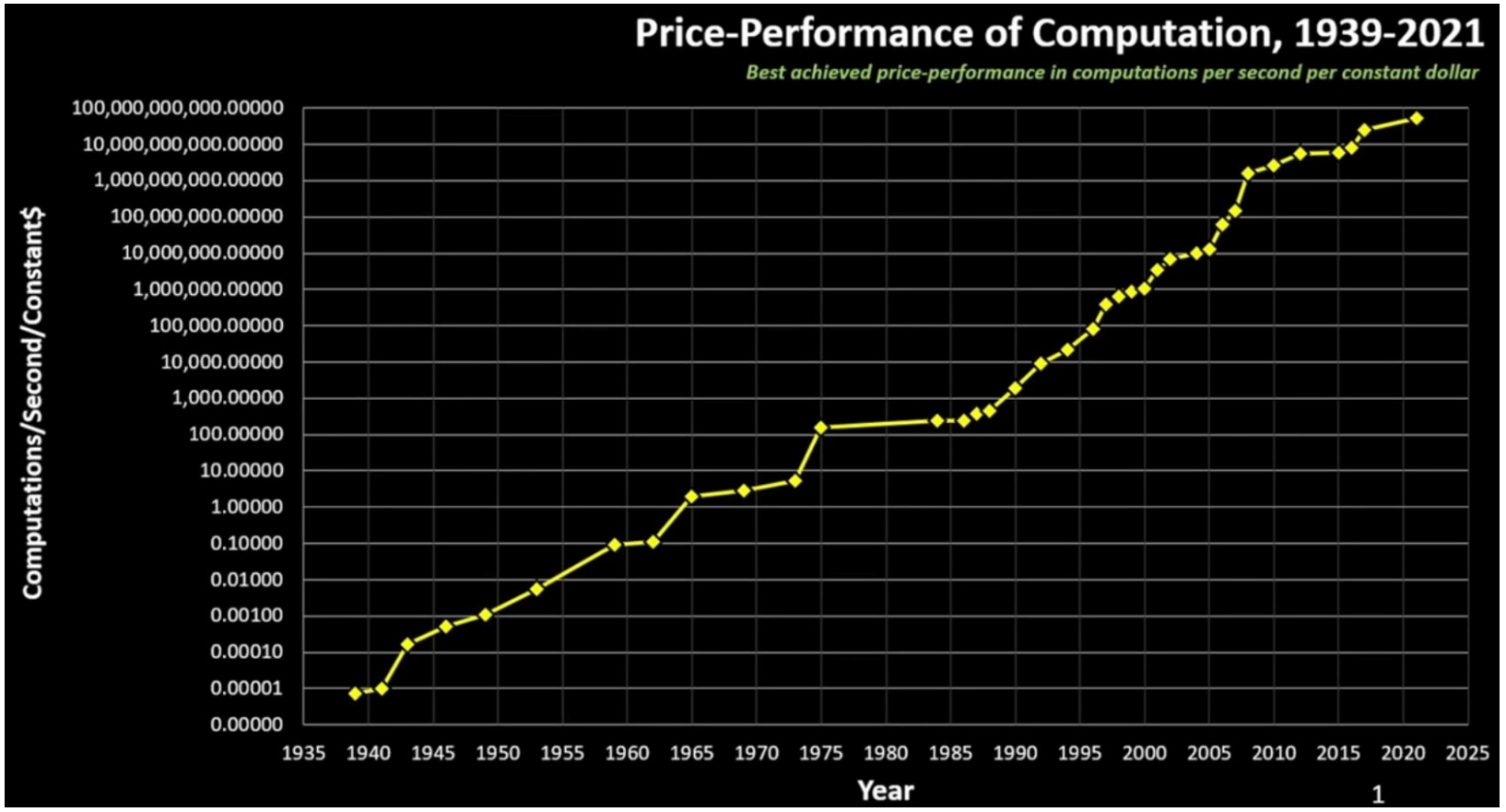
1965, Gordon Moore, founder of Intel:

„The number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years "... "for at least ten years"

- 10 µm – 1971
- 6 µm – 1974
- 3 µm – 1977
- 1.5 µm – 1982
- 1 µm – 1985
- 800 nm – 1989
- 600 nm – 1994
- 350 nm – 1995
- 250 nm – 1997
- 180 nm – 1999
- 130 nm – 2001
- 90 nm – 2004
- 65 nm – 2006
- 45 nm – 2008
- 32 nm – 2010
- 22 nm – 2012
- **14 nm – 2014**
- **10 nm – 2017**
- 7 nm – *~2019*
- **5 nm** – *~2021*

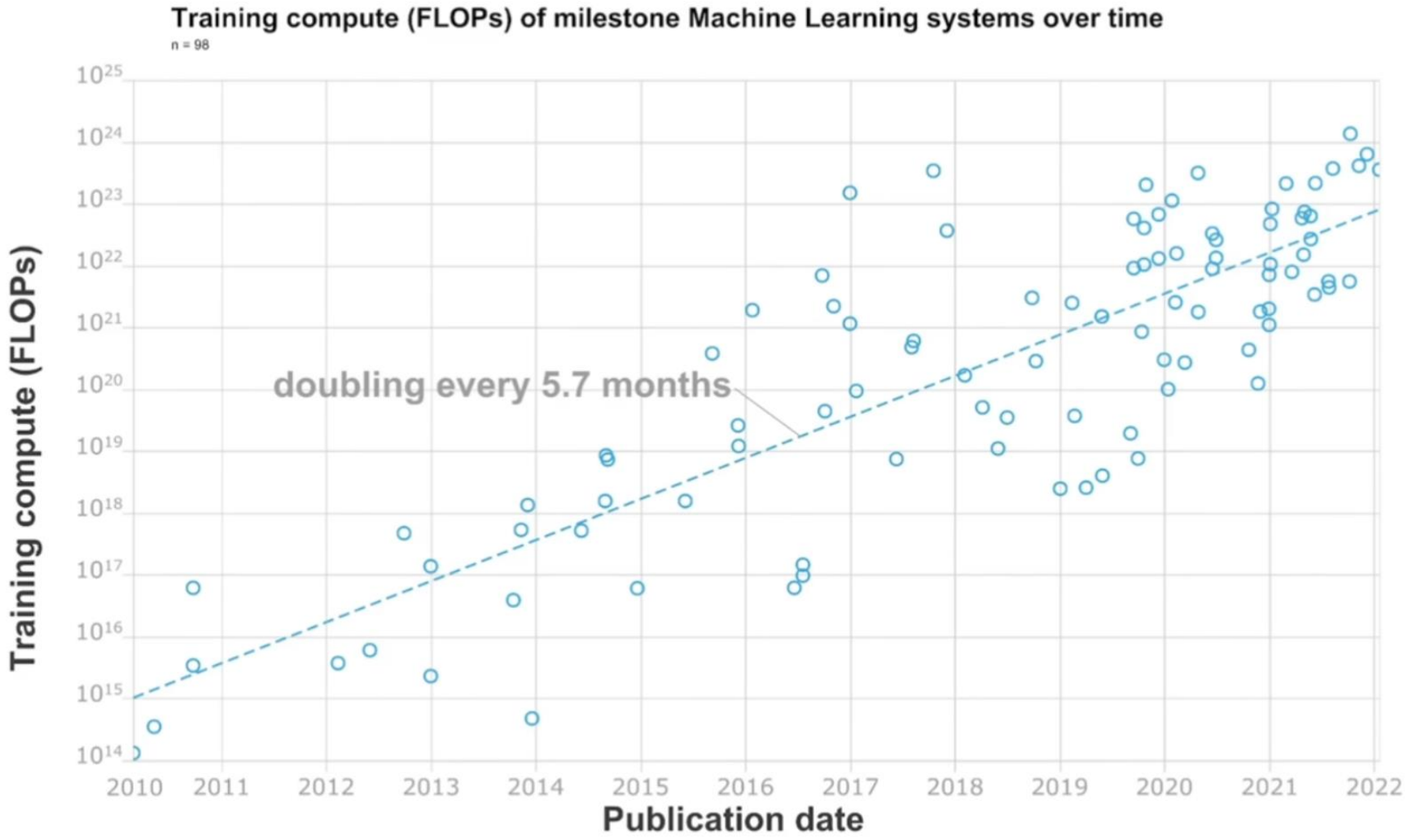
2012: single atom transistor (~0.1n, 1A)

A számítás egységnyi költségének alakulása



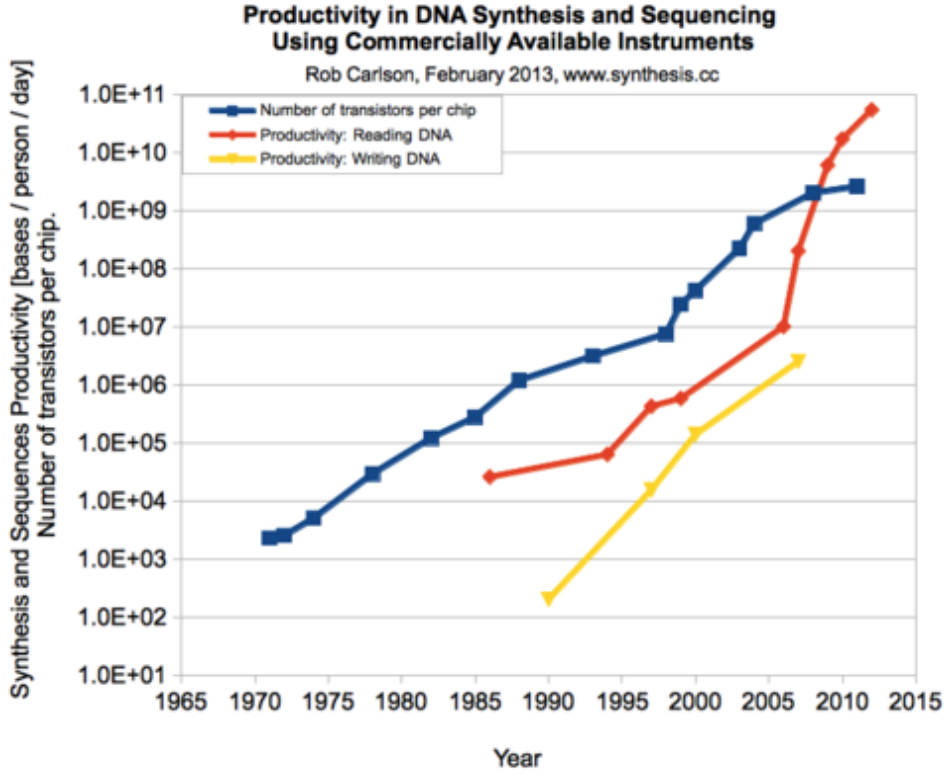
Ray Kurzweil: Singularity, Superintelligence, and Immortality (https://lexfridman.com/ray-kurzweil/)

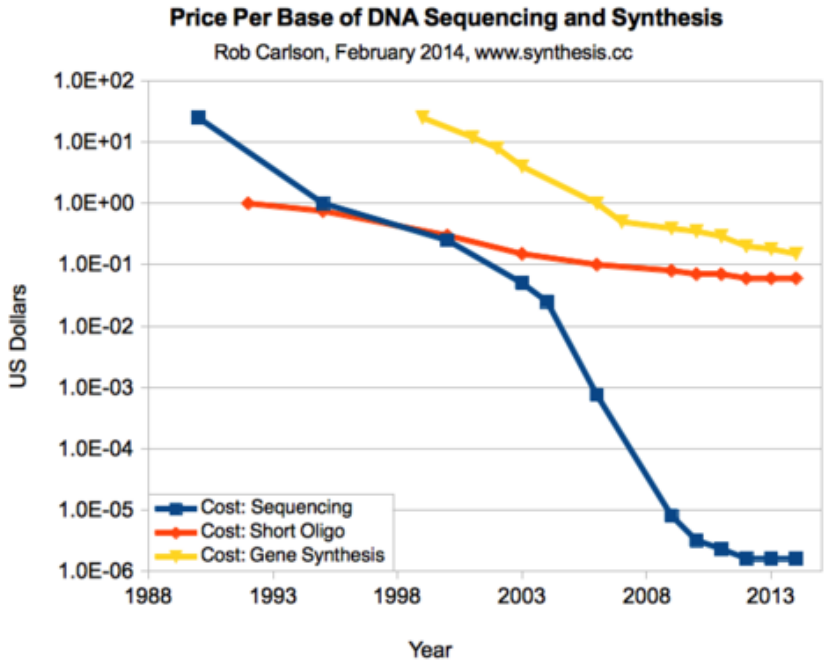
Gépi tanulásban használt számítás mennyisége



Ray Kurzweil: Singularity, Superintelligence, and Immortality (https://lexfridman.com/ray-kurzweil/)

Molekuláris biológia adatok költsége: Carlson's law





Number of biomedical publications

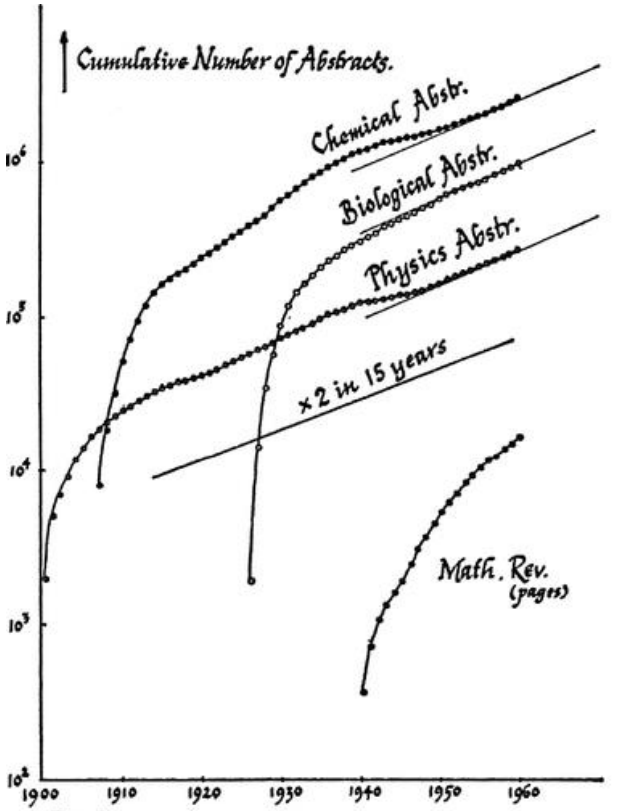
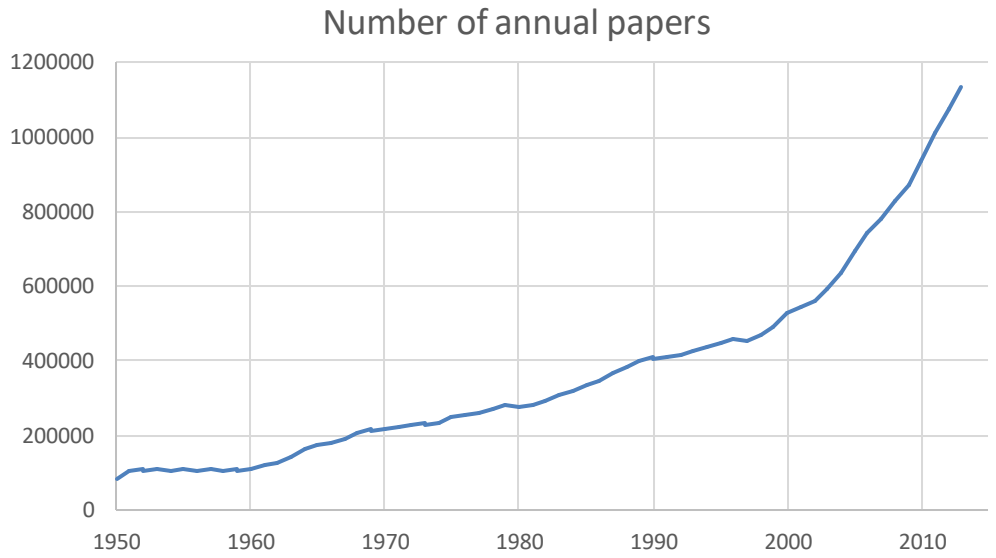


Fig. 2. CUMULATIVE NUMBER OF ABSTRACTS IN VARIOUS SCIENTIFIC FIELDS, FROM THE BEGINNING OF THE ABSTRACT SERVICE TO GIVEN DATE

It will be noted that after an initial period of rapid expansion to a stable growth rate, the number of abstracts increases exponentially, doubling in approximately 15 years.

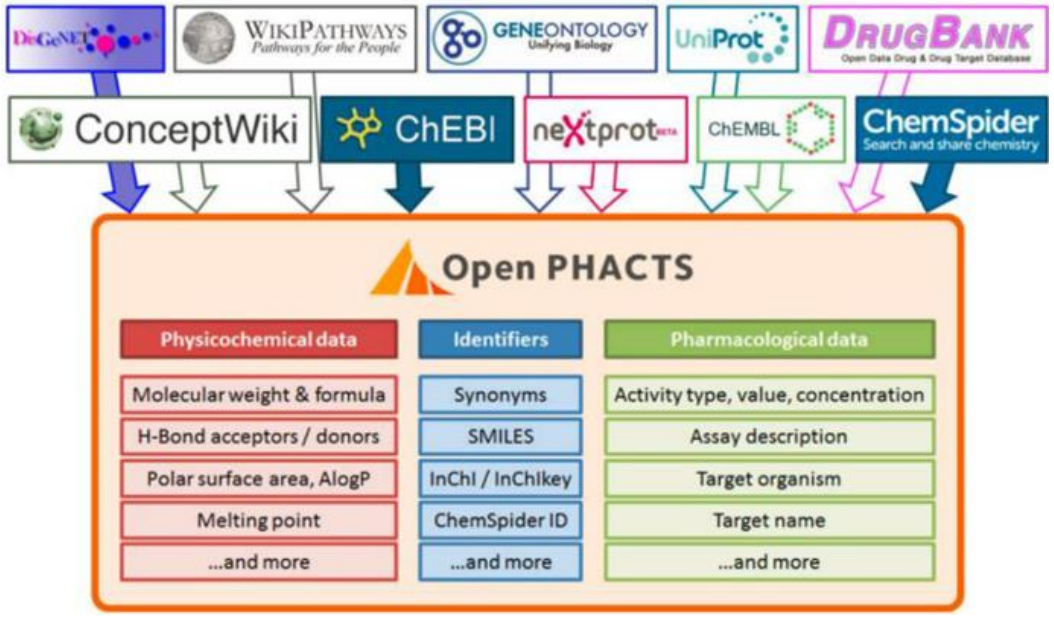
Little Science, Big Science, by Derek J. de Solla Price, 1963



Number of patents are in the same range...

Open knowledge

Semantic publishing



Williams, Antony J., et al. "Open PHACTS: **semantic interoperability** for drug discovery." Drug discovery today, 2012

Dumontier, Michel, et al. "Bio2RDF release 3: a larger connected network of **linked data** for the life sciences, EUR-WS, 2014.

[OPENBEL:]Hofmann-Apitius, Martin, et al. "Towards the taxonomy of human disease." Nature reviews. Drug discovery, 2015

M. Gerstein, "**E-publishing on the Web**: Promises, pitfalls, and payoffs for bioinformatics," *Bioinformatics,* 1999

M. Gerstein: **Blurring the boundaries between scientific 'papers' and biological databases**, Nature, 2001

P. Bourne, "Will a biological database be different from a biological journal?," *Plos Computational Biology,* 2005

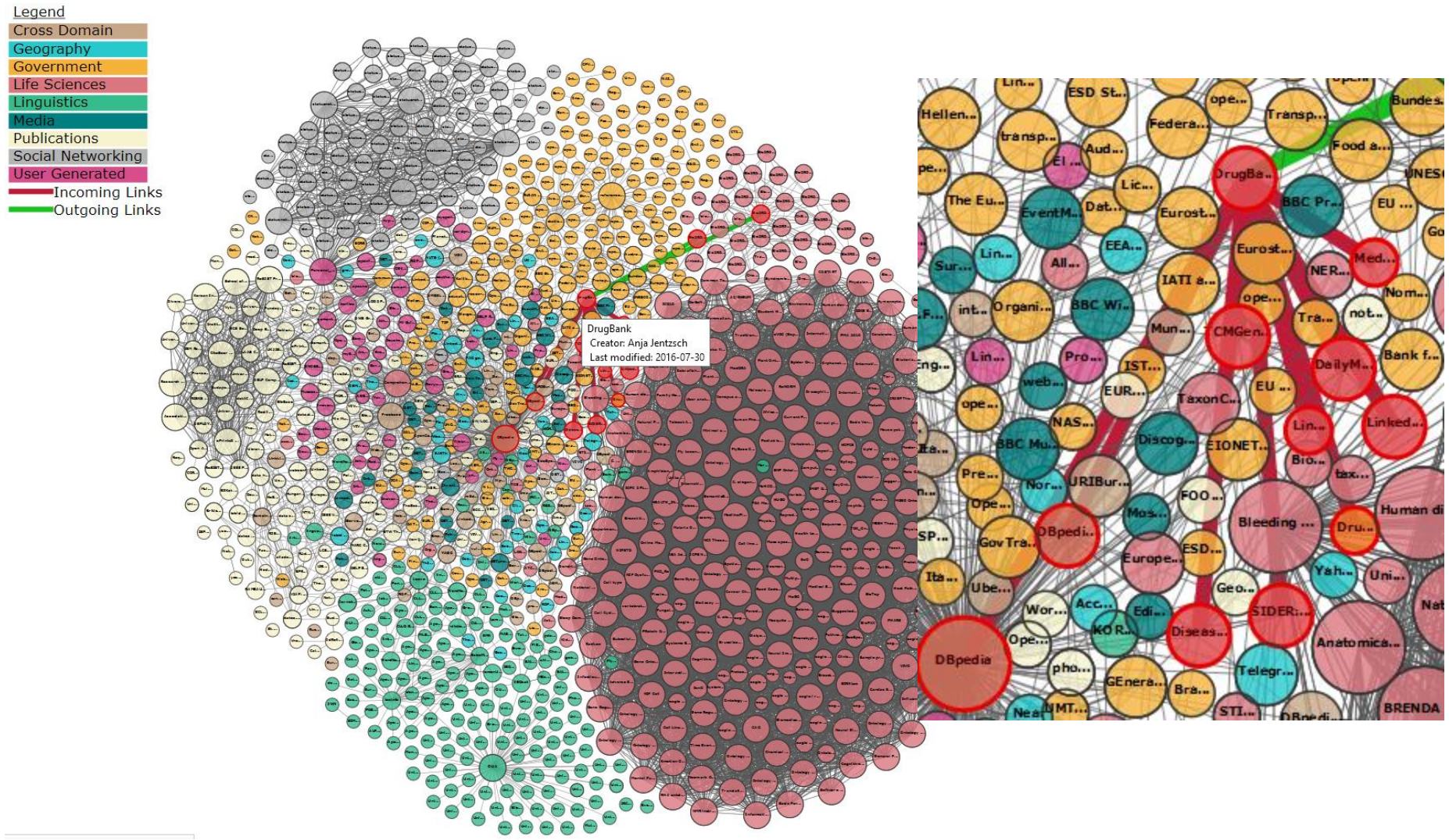
M. Gerstein et al: "**Structured digital abstract** makes text mining easy," *Nature,* 2007.

M. Seringhaus et al: "Publishing perishing? Towards tomorrow's information architecture," *Bmc Bioinformatics,* 2007.

M. Seringhaus: "Manually structured digital abstracts: A scaffold for automatic text mining," *Febs Letters,* 2008.

D. Shotton: "**Semantic publishing**: the coming revolution in scientific journal publishing," *Learned Publishing,* 2009

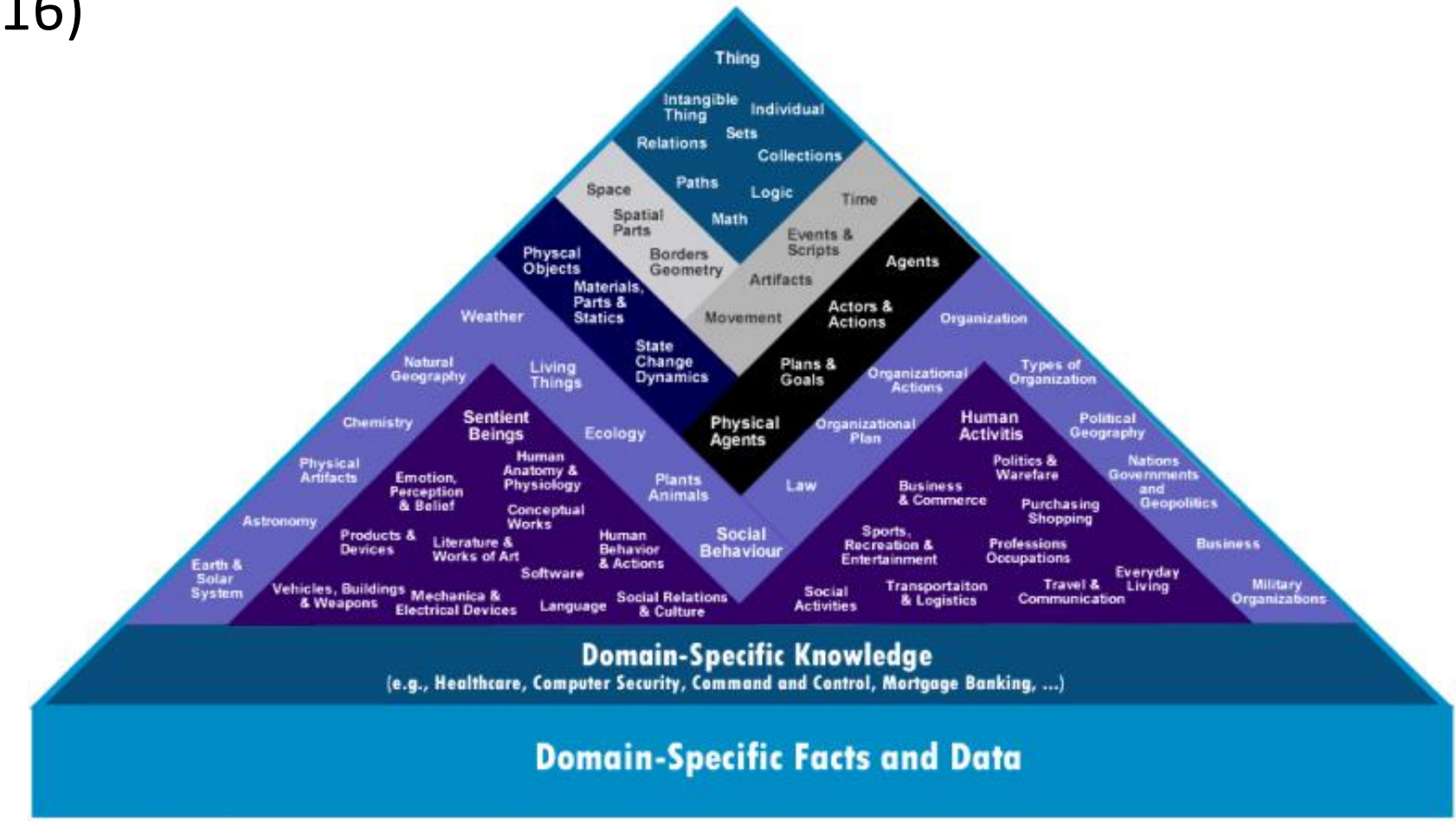
Knowledge: Linked open data



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

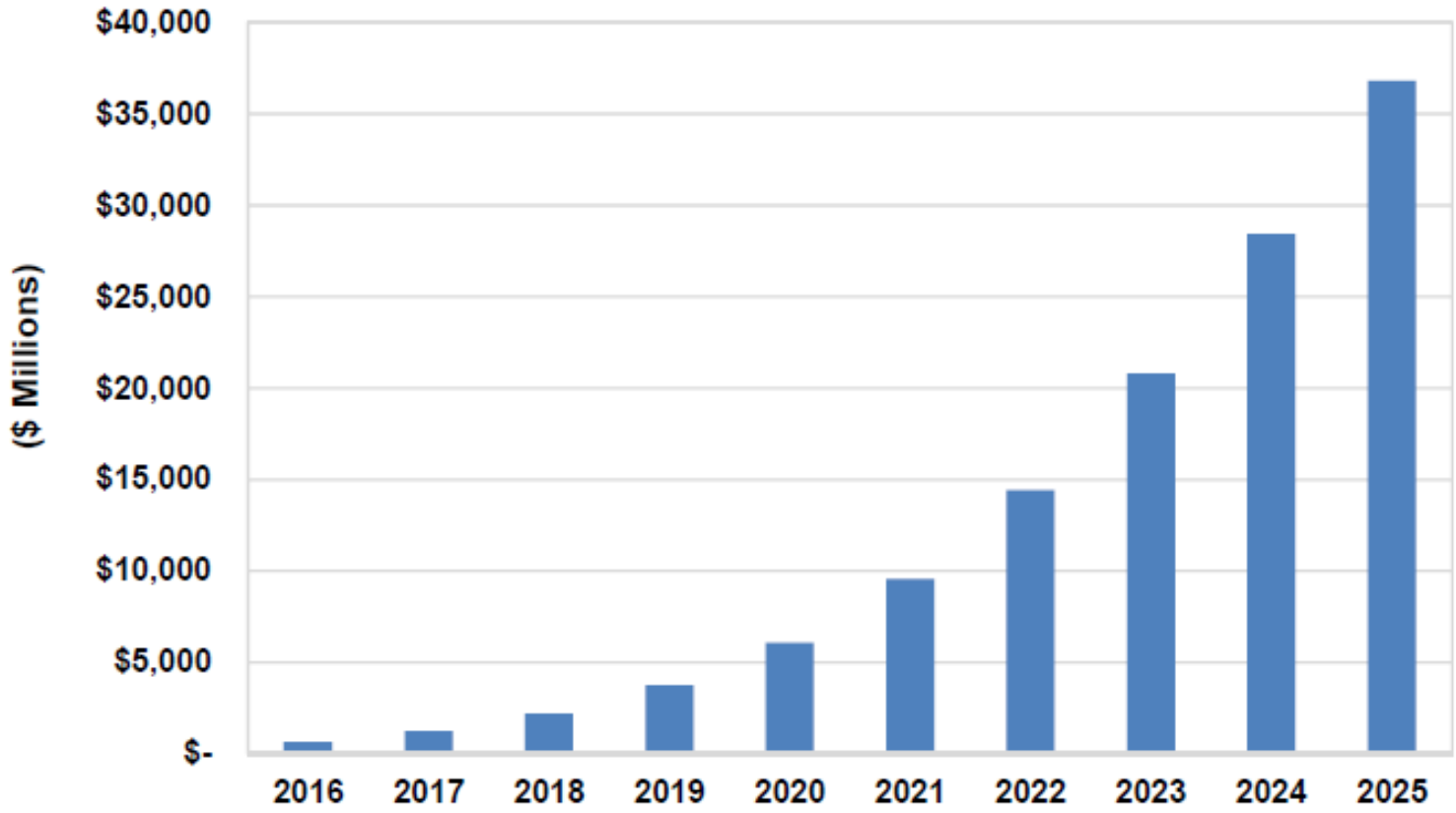
A józan ész szabályai

- The Cyc project (1984-2016)
- Goal: common sense
- Estimations in 1984:
  - 250 000 rules
  - 350 man-year
- Language: CycL
- Access: OpenCyc
- Current state
  - 239,000 concept
  - 2,093,000 facts



MI piacok & pénzügyi források

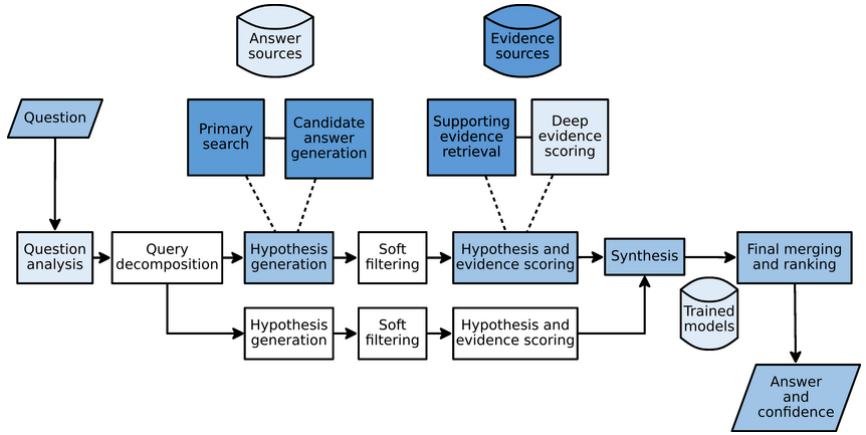
Chart 1.1 Artificial Intelligence Revenue, World Markets: 2016-2025



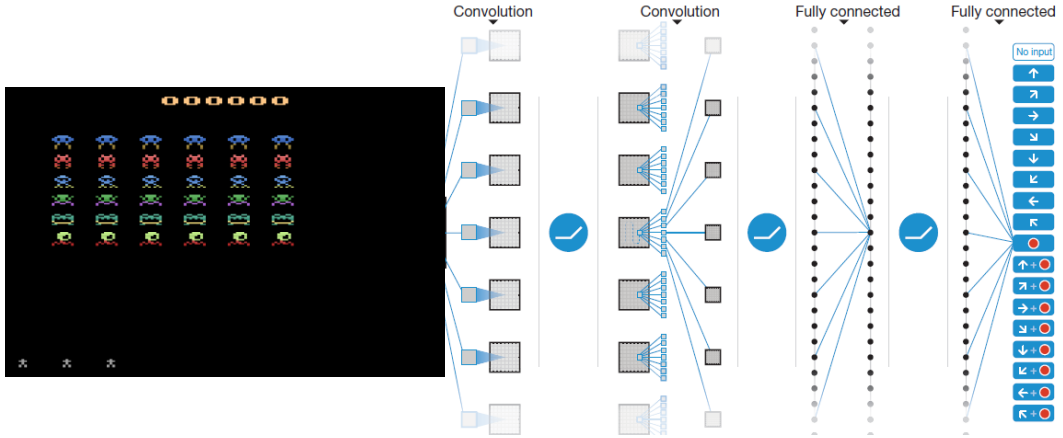
(Source: Tractica)

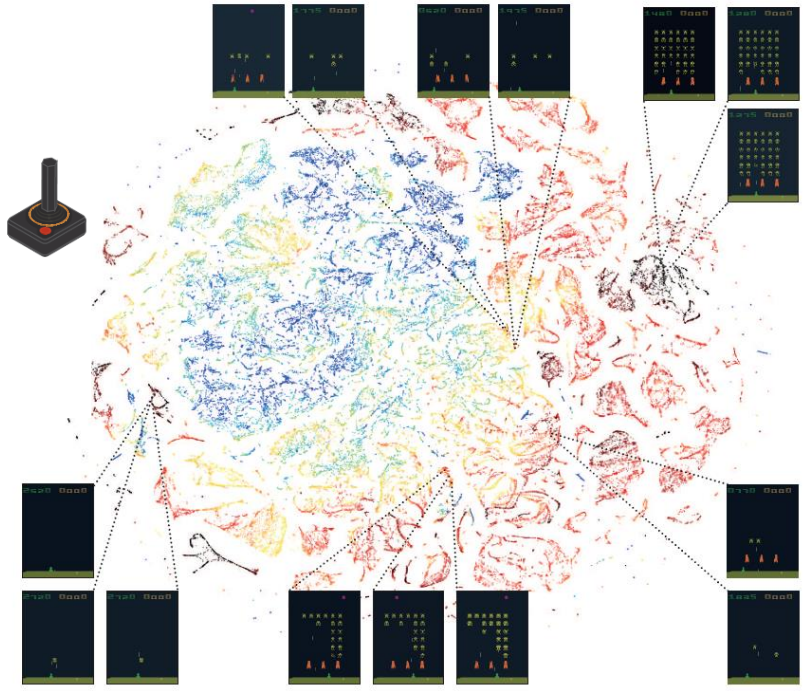
IBM Watson (2011): Jeopardy

- IBM Grand Challenge
  - 1997: **Deep Blue** wins human champion G. Kasparov.
  - 1999-2006<: **Blue Gene**, protein prediction
  - 2011: **Watson**
    - Natural language processing
    - inference
    - Game theory



Playing computer games (2015)





LETTER

doi:10.1038/nature14236

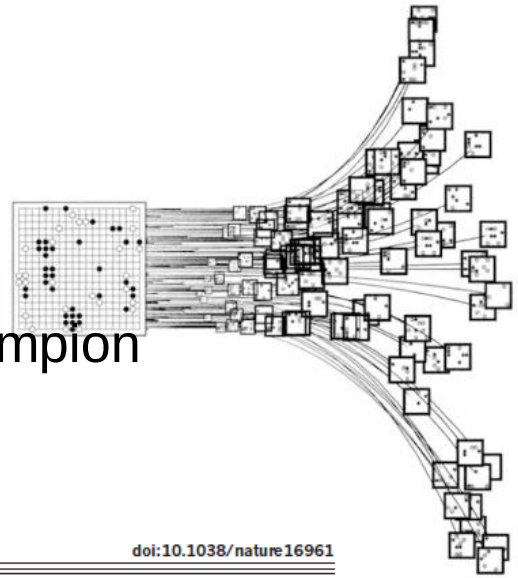
Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

Go (2017)

AlphaGo

- Google DeepMind
- Monte Carlo tree search
- 2016: 9 dan
- 2017: wins against human champion



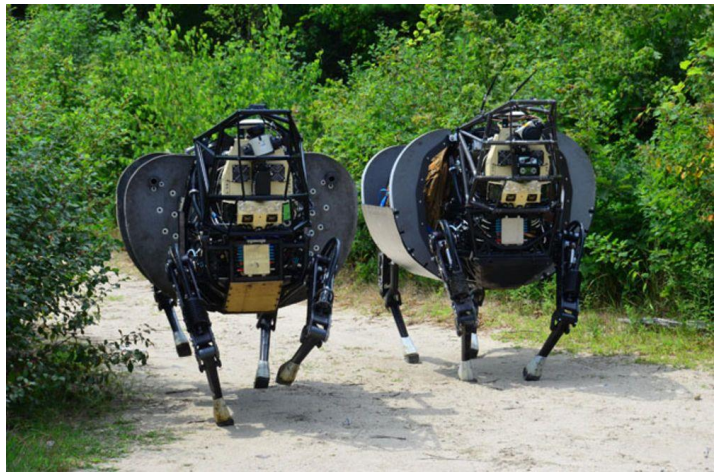
ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

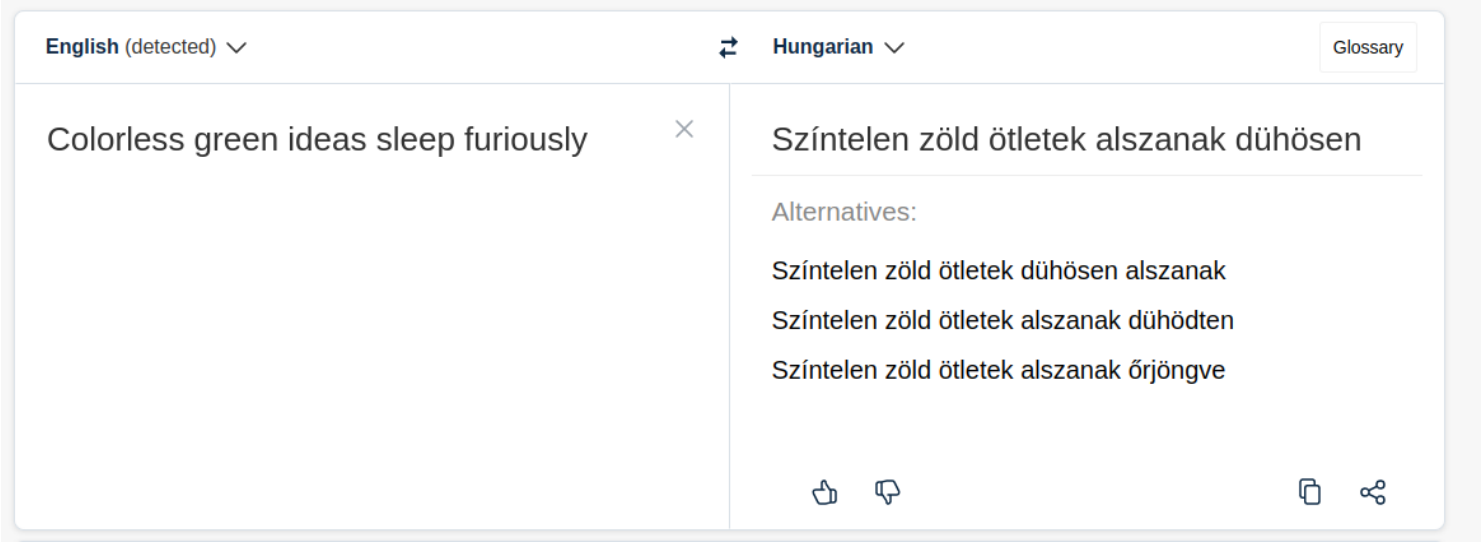
David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

Walking, movements

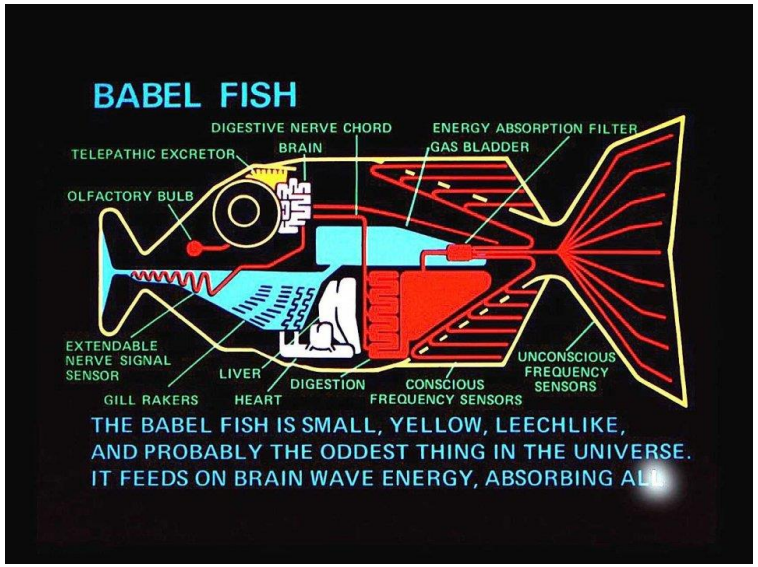


(Real-time) translation

Pilot Translating Earpiece

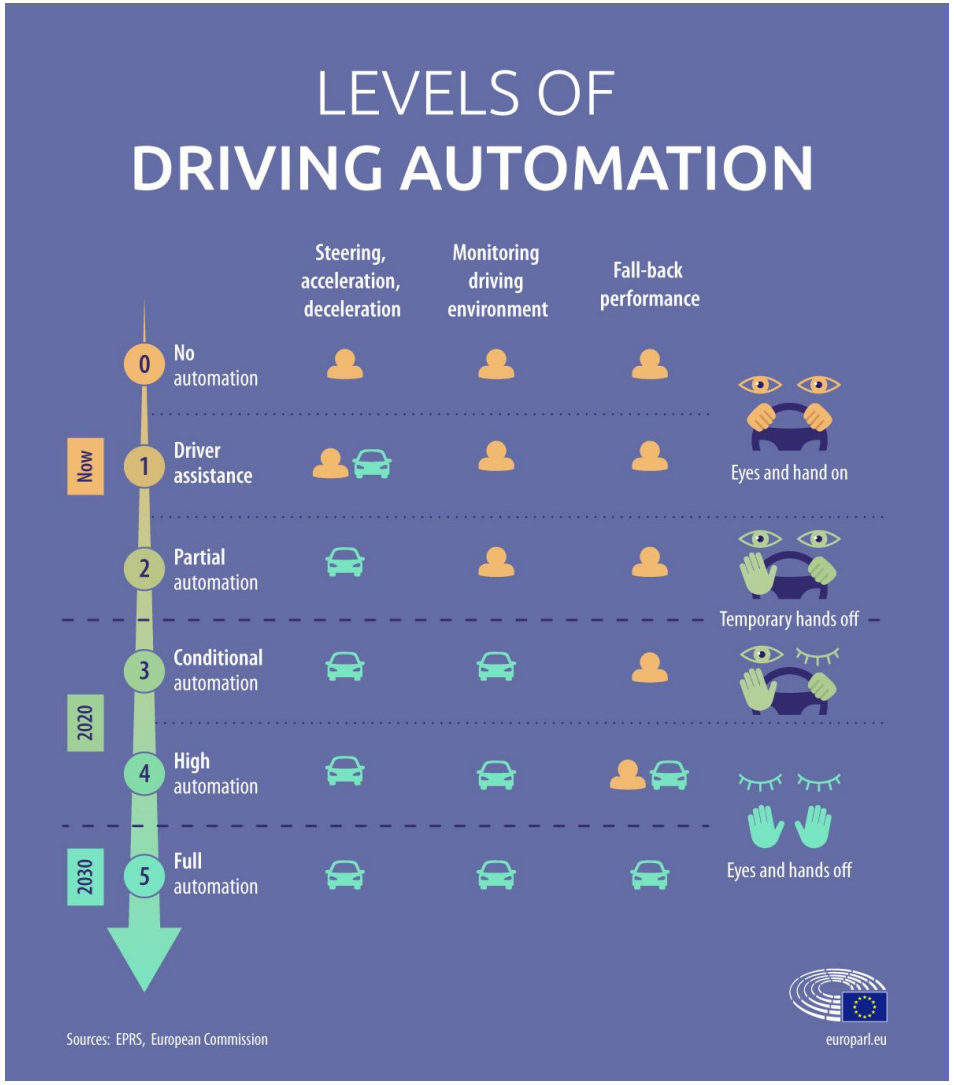


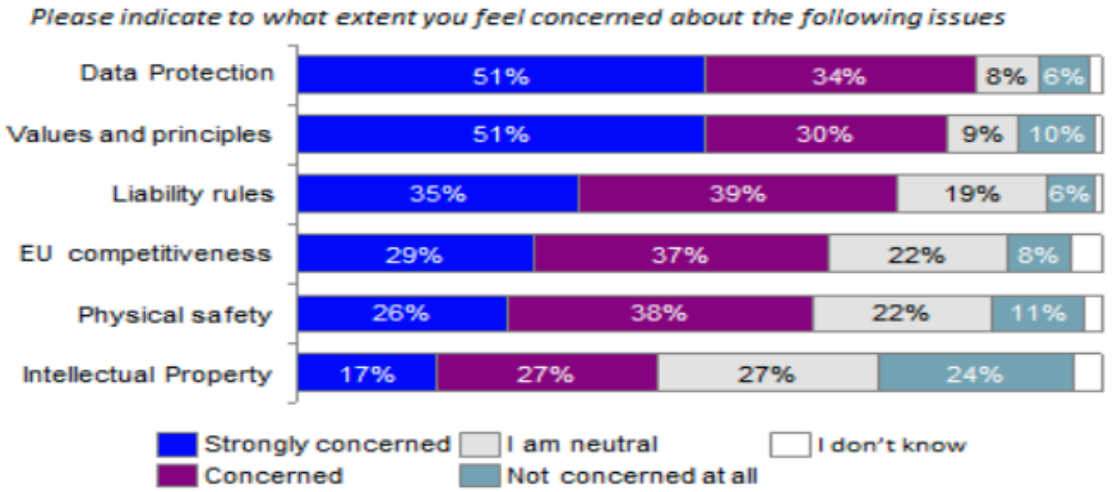
https://www.deepl.com/translator

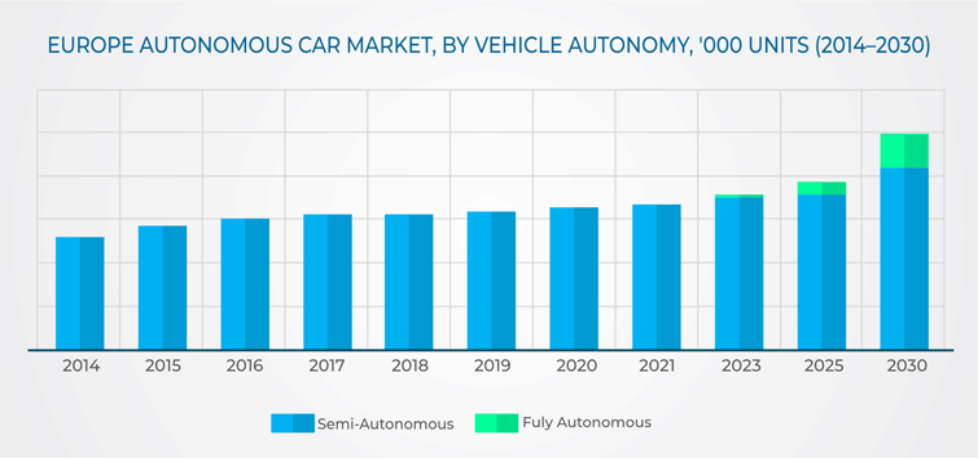


D.Adams: Galaxis útikalauz stoppososoknak
Hitchhiker's Guide to the Galaxy"

Autonomous driving

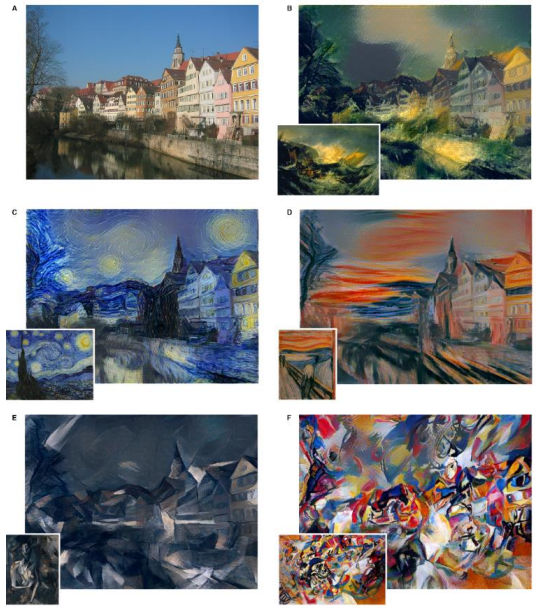


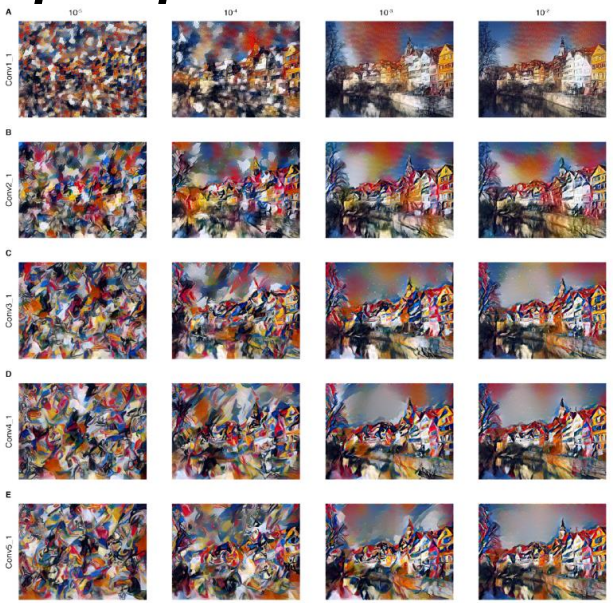




Artistic style

- Gatys, L.A., Ecker, A.S. and Bethge, M., 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.





- Stable diffusion

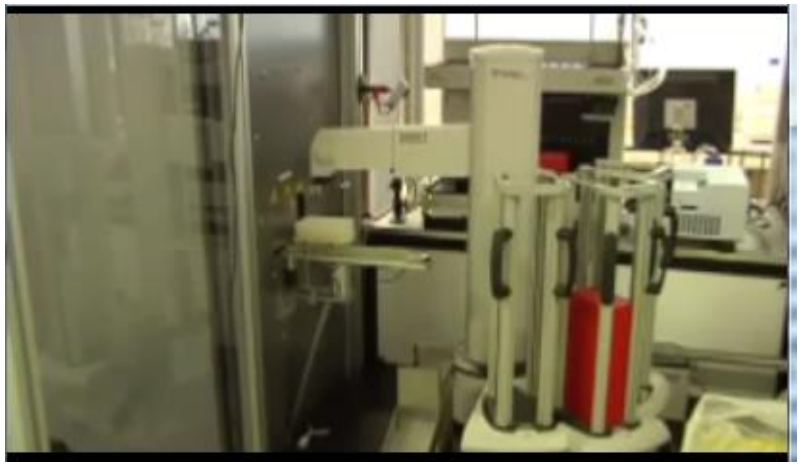
  https://huggingface.co/spaces/stabilityai/stable-diffusion

Digital arts and design

https://edition.cnn.com/2022/09/03/tech/ai-art-fair-winner-controversy/index.html
https://edition.cnn.com/style/article/ai-architecture-manas-bhatia/index.html
https://www.midjourney.com/home/

Automated discovery

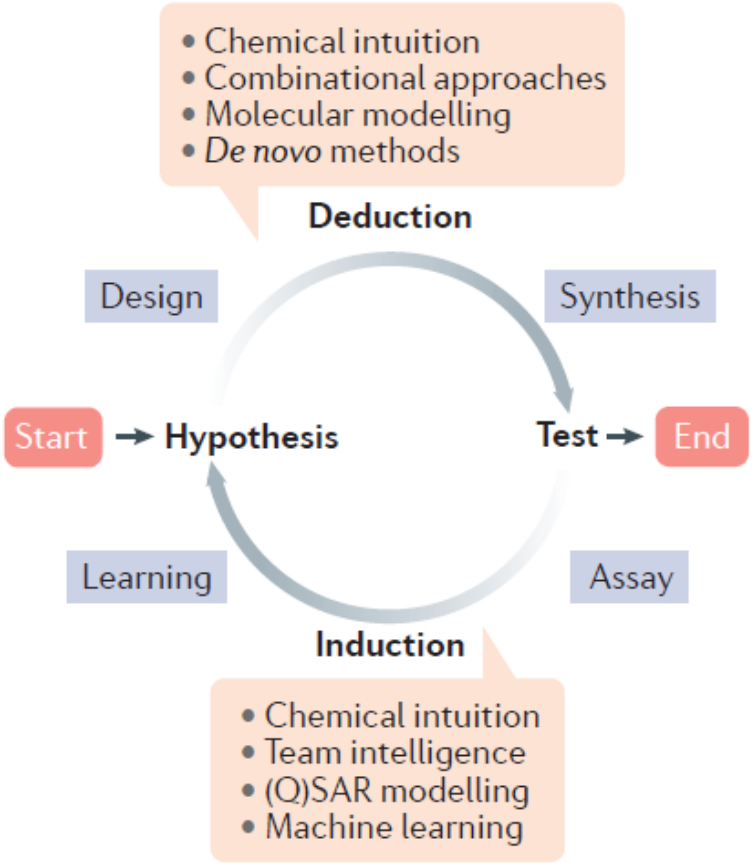
- Langley, P. (1978). Bacon: A general discovery system.
- …
- …
- R.D.King et al.: **The Automation of Science**, Science, 2009
- Sparkes, Andrew, et al.: Towards **Robot Scientists** for autonomous scientific discovery, 2010



„Adam”

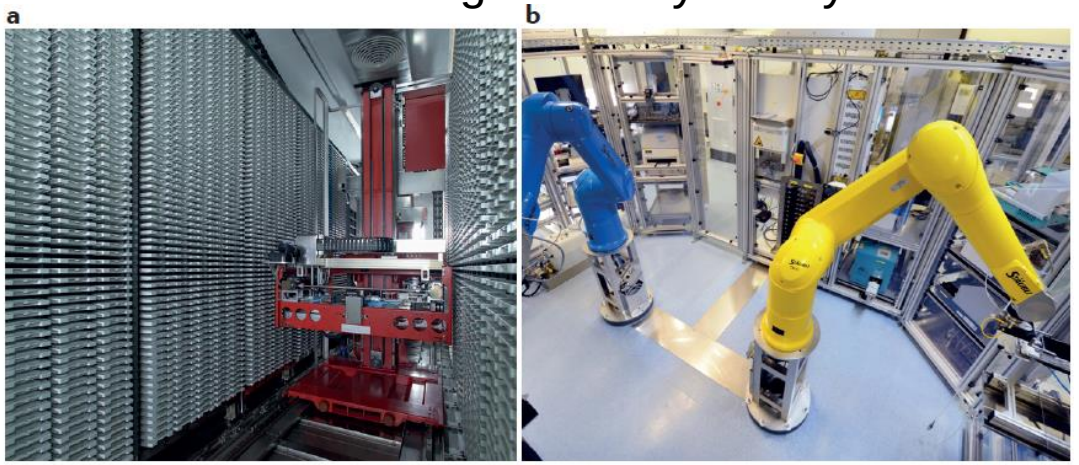
„Eve”

Automating drug discovery

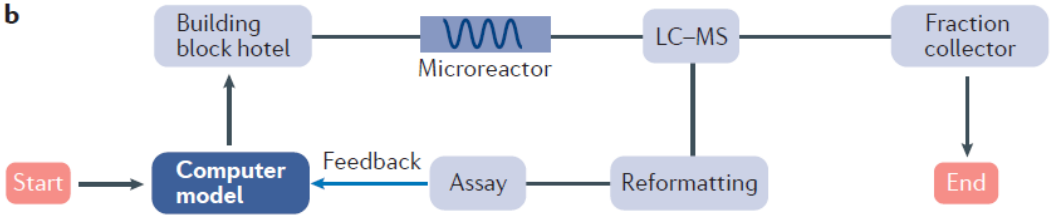


Design cycle

Automated drug discovery facility



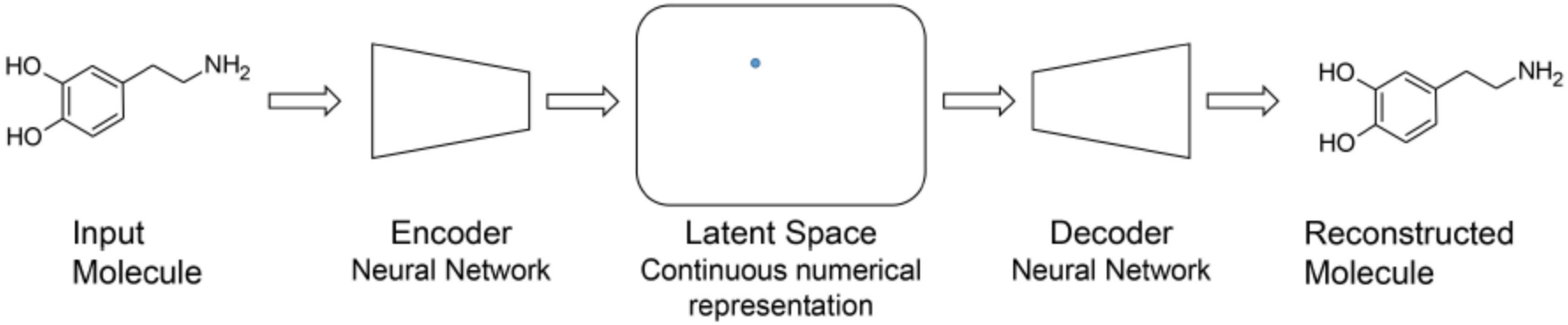
Active learning with microfluidics



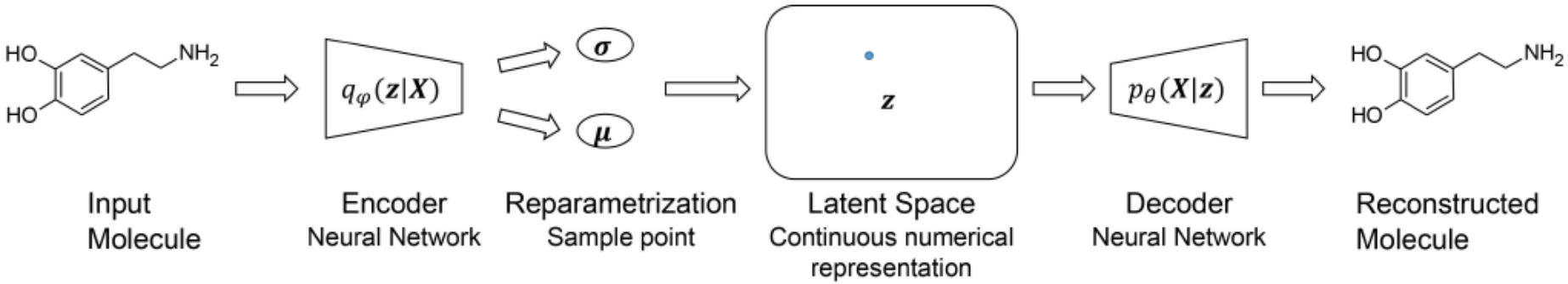
Schneider, Gisbert. "Automating drug discovery." *Nature Reviews Drug Discovery* 17.2 (2018): 97.

De novo molecular design

Autoencoder

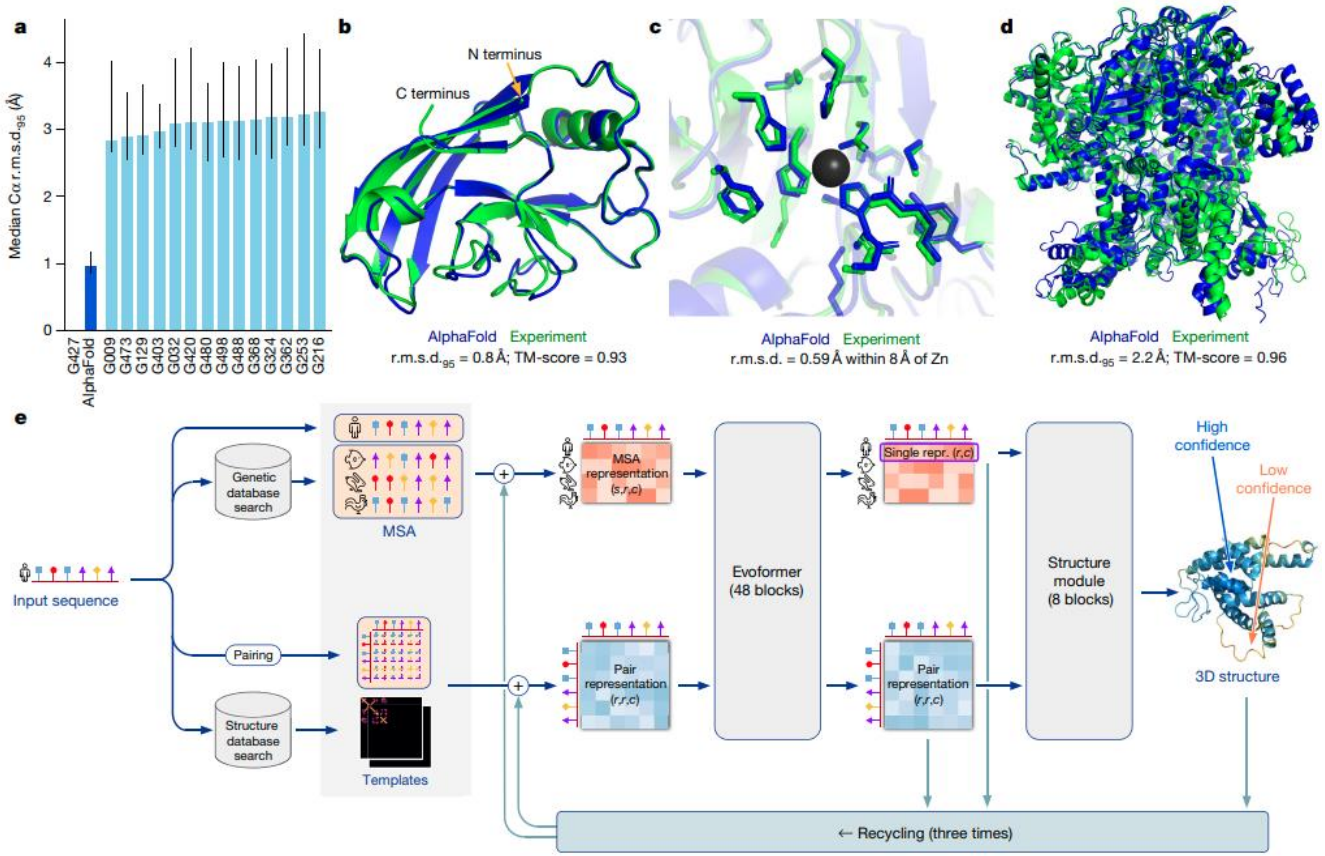


Variational autoencoder



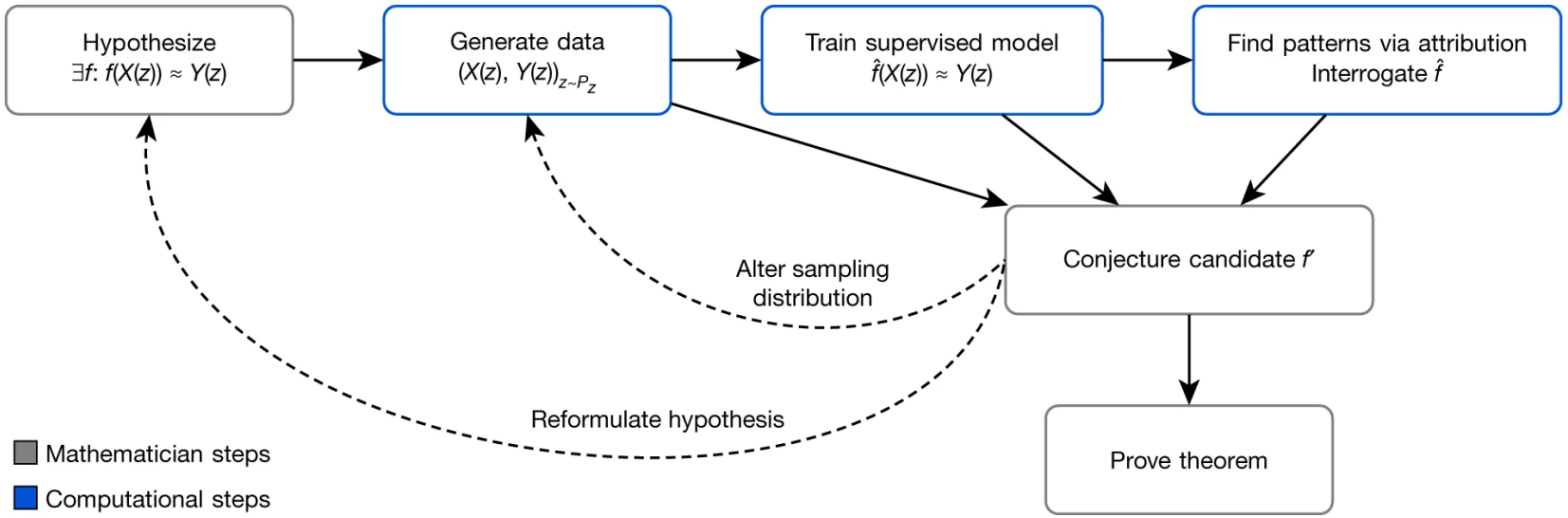
Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. and Chen, H., 2018. Application of generative autoencoder in de novo molecular design. *Molecular informatics*, *37*(1-2), p.1700123.

AlphaFold (2020)



Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.

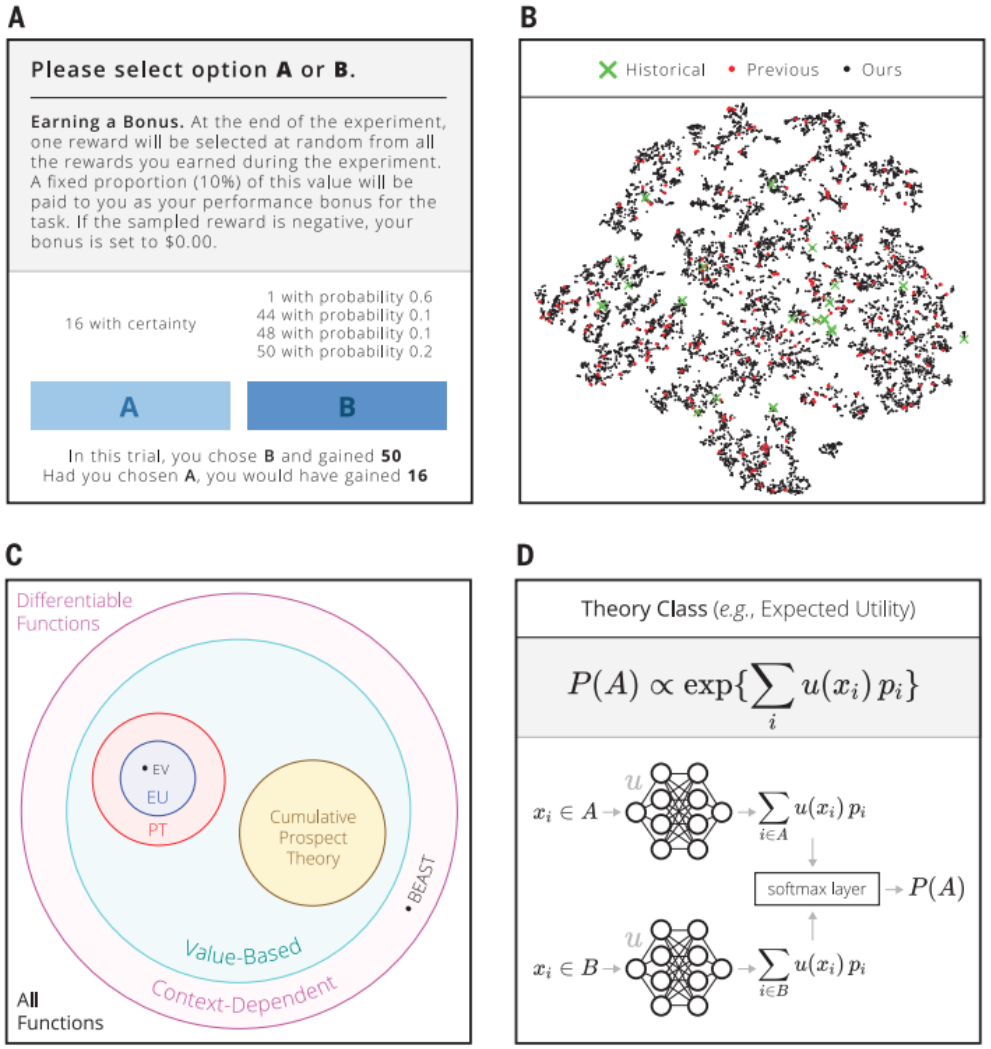
Mathematical discovery/creativity



z: **Knot**	$X(z)$: **Geometric invariants**				$Y(z)$: **Algebraic invariants**		
	Volume	Chern–Simons	Meridional translation	...	Signature	Jones polynomial	...
	2.0299	0	i	...	0	$t^{-2} - t^{-1} + 1 - t + t^2$	...
	2.8281	−0.1532	$0.7381 + 0.8831i$	...	−2	$t - t^2 + 2t^3 - t^4 + t^5 - t^6$	...
	3.1640	0.1560	$-0.7237 + 1.0160i$	...	0	$t^{-2} - t^{-1} + 2 - 2t + t^2 - t^3 + t^4$	...

Davies, Alex, et al. "Advancing mathematics by guiding human intuition with AI." *Nature* 600.7887 (2021): 70-74.

Black-box modeling of human judgement



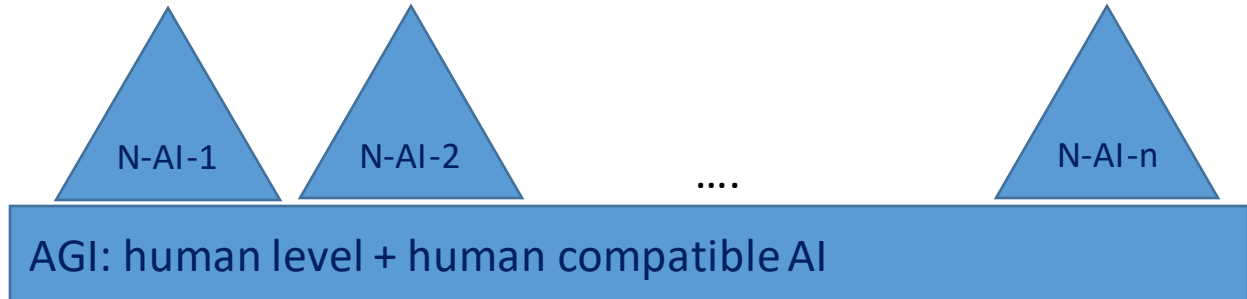
Peterson, Joshua C., et al. "Using large-scale experiments and machine learning to discover theories of human decision-making." *Science* 372.6547 (2021): 1209-1214.

Szűk, széles és általános/emberszintű MI

- Diverse AI solutions
  - Vision, robotics, natural language processing (NLP),…
  - Recommendation systems: what to read, buy, listen,...
  - Industry 4.0: planning, production, testing, logistics
  - Self-driving cars
  - Automated scientific discovery systems

- AGI tests
  - The Turing Test (Turing)
  - The Coffee Test (Wozniak)
  - The Robot College Student Test (Goertzel)
  - The Employment Test (Nilsson)
  - The flat pack furniture test (Severyns)
  - …

- Unification of AGI
  - Unified definition
  - Unified measure
  - Unified framework



Források emberszintű MI-hez

- AGI Society (Artificial General Intelligence, AGI)
  - http://www.agi-society.org/
- AGI conferences 2008..
  - https://en.wikipedia.org/wiki/Conference_on_Artificial_General_Intelligence
  - 2008: AGI-08 Workshop on the Sociocultural, Ethical and Futurological Implications of Artificial General Intelligence
    - http://agi-conf.org/2008/workshop/
  - http://agi-conf.org/2019/

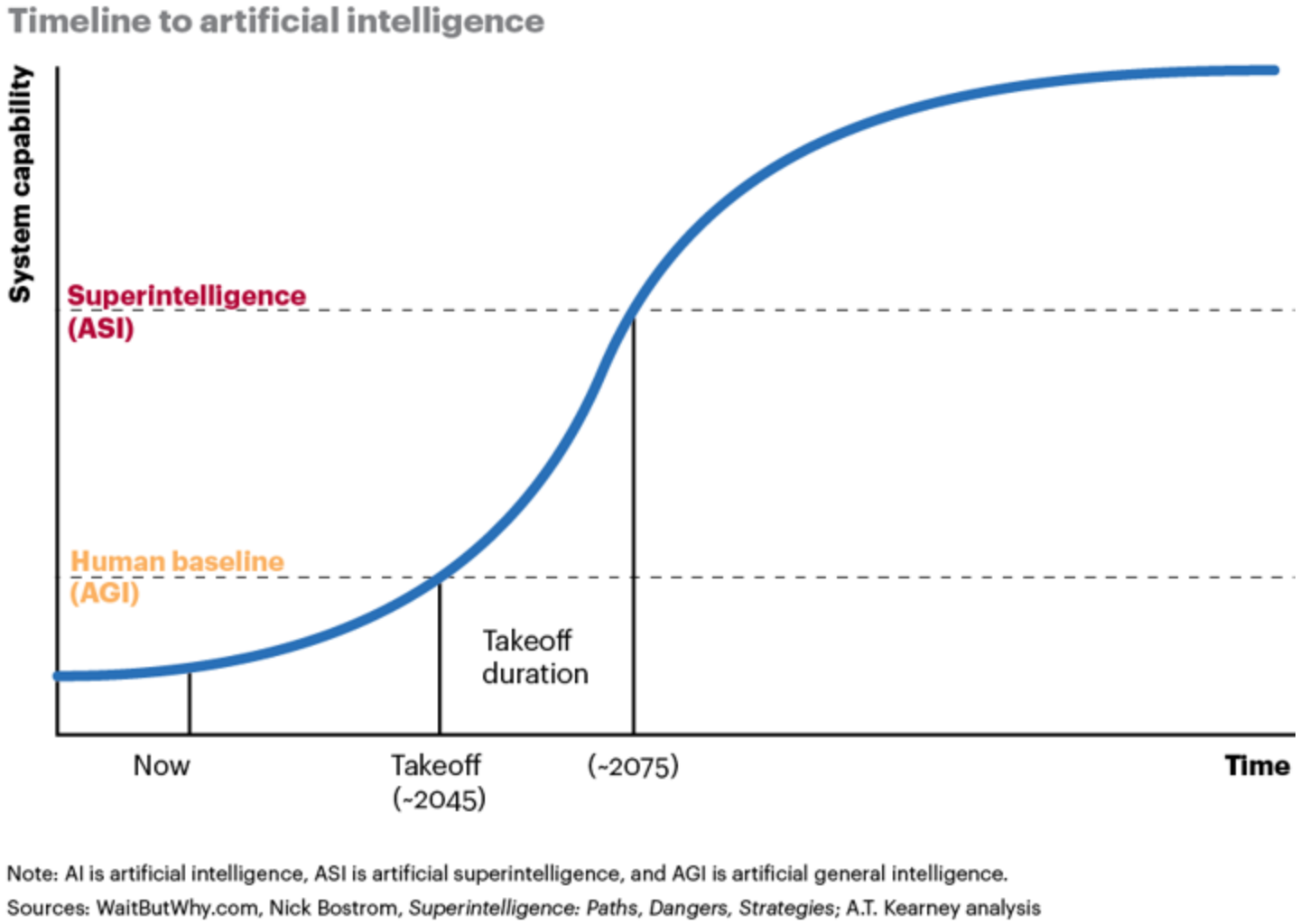
Intelligence explosion

„Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an **'intelligence explosion**,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control." Good (1965),

Superintelligence and technological singularity

Expert consensual prediction for date of Turing-test:

- 1999:
  - 20%: never
  - 80%: 100y
- 2020:
  - 2042
- 2020:
  - 2030



Kurzweil, Ray. *The singularity is near: When humans transcend biology*. Penguin, 2005.
Nick Bostrom: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford, 2014

* (?)

Az MI* ígérete

2045

Az MI veszélyei

A "gorilla" probléma

Russell, Stuart: New AI, AI25@BME VIK, 2018

Russell, Stuart. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

Egzisztenciális veszélyek az emberiségre

Az emberiség kihalásával (földi élet...modern civilizáció eltűnésével) fenyegető egzisztenciális veszélyek

- Kozmikus katasztrófa (aszteroida, napkitörés,..)
- Klímaváltozás
- Erőforrások kimerülése
- Nukleáris háború
- szuperMI?

Superintelligence

- A.Turing(1951): "turning off the power"
- N.Wiener (1964): "In the past, a partial and inadequate view of human purpose has been relatively innocuous only because it has been accompanied by technical limitations."
- S.Hawking, Bill Gates, Elon Musk… (2015): **Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter**

Az MI* egzisztenciális veszélye: ellenérvek

Sokfélék

- Nem aktuális/rémisztgetés/szenzációhajhászás
- Nem felügyelt tanulás megszokottak az MI-ben
- Önjavító/boosting MI technikák elterjedtek
- Standard mérnöki biztonsági keretek kezelik
- Előnyök-nagyobbak, mint hátrányok
- Jobb elsőnek magunknak létrehozni
- Lekorlátozhatóság
- Kikapcsolási lehetőség (Switch-off)

Russell, Stuart. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

Már megvalósult MI veszélyek

1. Politikai befolyásolás
2. Információs buborék
3. Valószerű hamis hírek/információ (deep fake)
4. Társadalmi polarizáltság
5. Digitális függések
6. Mentális betegségek (testképzavarok)
7. Jövedelmi egyenlőtlenségek
8. Ökológiai lábnyom
9. Változó (leromló?) kognitív képességek
10. ….

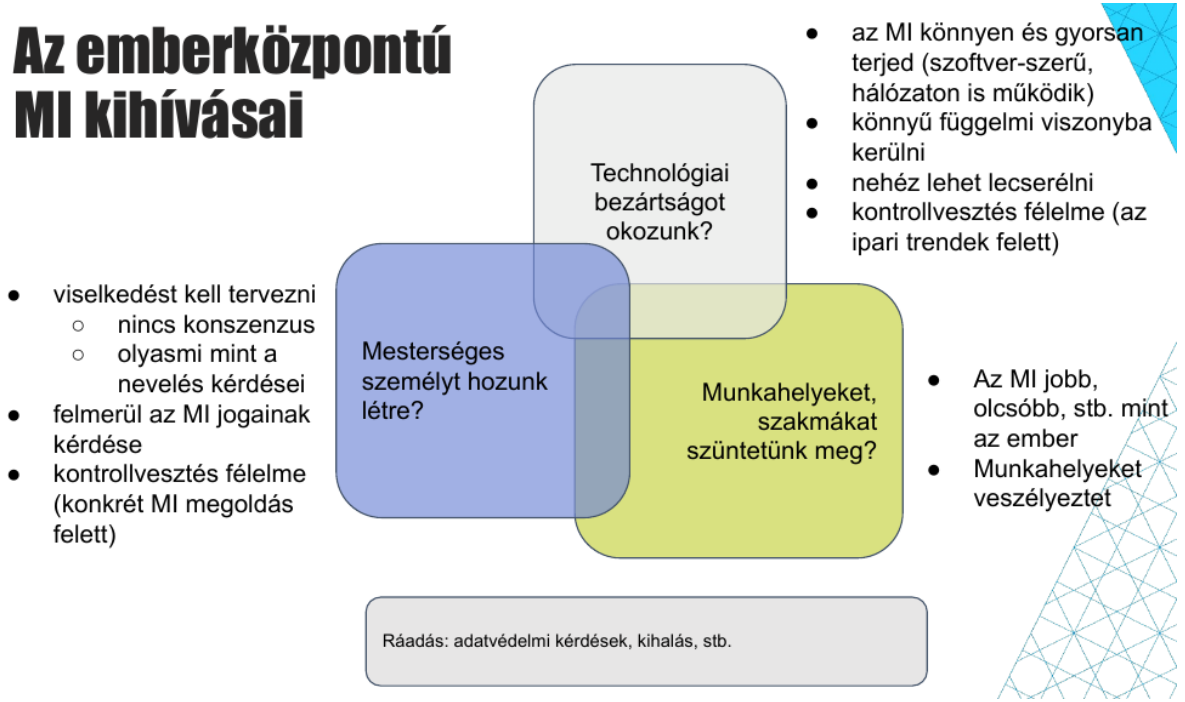
Digitális közösségi platformok:

- Marshall Mcluhan (1964): "The medium is the message"
- **Social Dilemma** (2020)
  - "What I want people to know is that everything they're doing online is being watched, is being tracked, is being measured" (Former Executive of Twitter, Jeff Seibert)
  - "It's not about the technology being the existential threat. It's the technology's ability to **bring out the worst** in society being the existential threat. (Tristan Harris)
- Nem javasol, hanem befolyásol

==> Az MI technológiák már egzisztenciális veszélyt jelentenek! (szuperMI elérése nélkül is)

Az MI és EMMI etikai kihívásai

- Lethal autonomous weapons
- Privacy: surveillance (cameras), cybersecurity
- Fairness and bias, fairness principles
- Polarization and segmentation in the society (information bubble)
- Fake news, deep fake
- Digital addiction and AI (internet, gaming,...)
- Social credit systems
- Trust and transparency
- Human compatible AI (on values)
- AI Safety
- The future of work
- Robot rights, patents



Russell, S. J., and Peter Norvig. "Artificial Intelligence: A Modern Approach, 4th, Global ed." (2022).

Héder Mihály: Az emberközpontú mesterséges intelligencia etikai kérdései (HTE 2022. szeptember 29.)

Elvek egy etikus (EM)MI-hez

- Ensure safety
- Establish accountability
- Ensure fairness
- Uphold human rights and values
- Respect privacy
- Reflect diversity/inclusion
- Promote collaboration
- Avoid concentration of power
- Provide transparency
- Acknowledge legal/policy implications
- Limit harmful uses of AI
- Contemplate implications for employment

Emberiesség: "deference to humans"
Értelmezhetőség: interpretability/understandability
Magyarázhatóság: explainability
Federáltság: autonomous collective intelligence

Russell, S. J., and Peter Norvig. "Artificial Intelligence: A Modern Approach, 4th, Global ed." (2022).

Emberközpontú MI: az emberi értékekhez igazodó MI

Új MI definíció egy bizonyíthatóan jóravaló MI-hez

- Az emberek olyan mértékben intelligensek, amennyire cselekedeteik várhatóan elérik céljaikat.
- A gépek olyan mértékben intelligensek, amennyire cselekedeteik várhatóan elérik céljaikat.
- **A gépek olyan mértékben jóravalóak, amennyire cselekedeteik várhatóan elérik céljainkat.**

Russell, Stuart. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

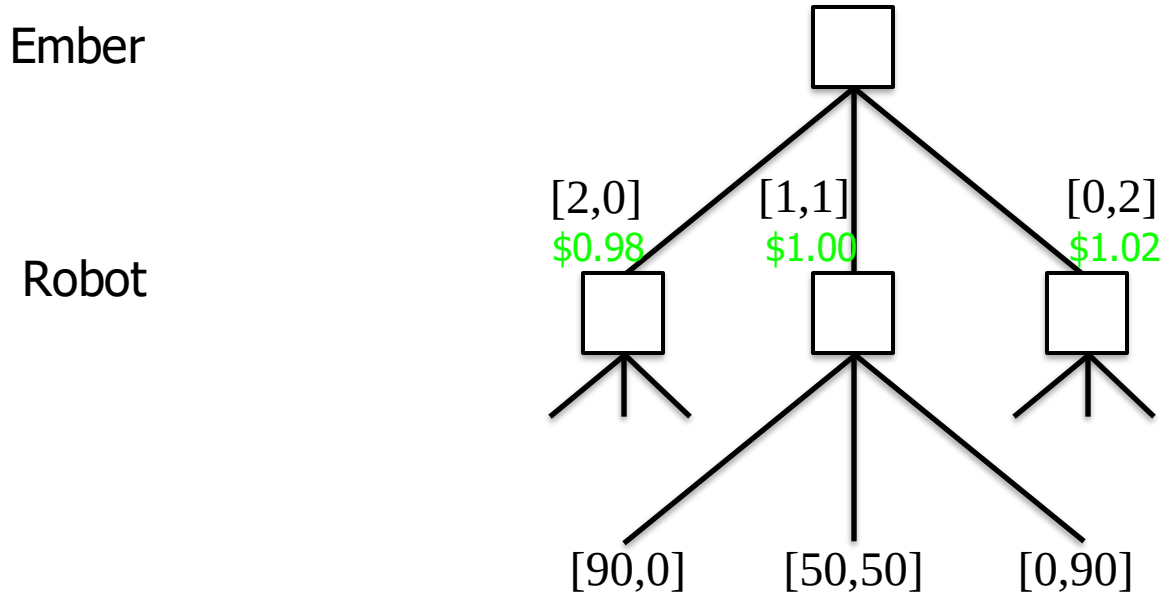
Három alapelv az emberhez igazodó MI-hez

1. A robot egyetlen célja, hogy maximalizálja az emberi preferenciák megvalósulását.
2. A robot kezdetben nem tudja, hogy mik ezek a preferenciák.
3. Az emberi preferenciákról a legjobb információforrás az emberi viselkedés

Russell, Stuart. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

EMMI mint Nash-egyensúly

- Egy (p,s) állapotban p gemkapocs és s kapocs van.
- Az emberi jutalom θp + (1-θ)s és θ=0,49
- A robotnak egyenletes priorja van a θ értékre a [0,1]-en.

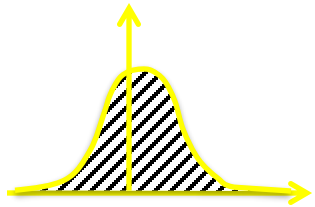


[1,1] optimális

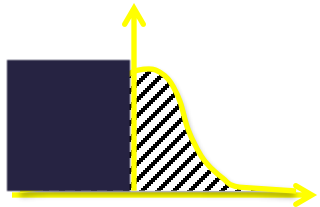
Russell, Stuart: New AI, AI25@BME VIK, 2018
Russell, Stuart. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

Az "off-switch" modell

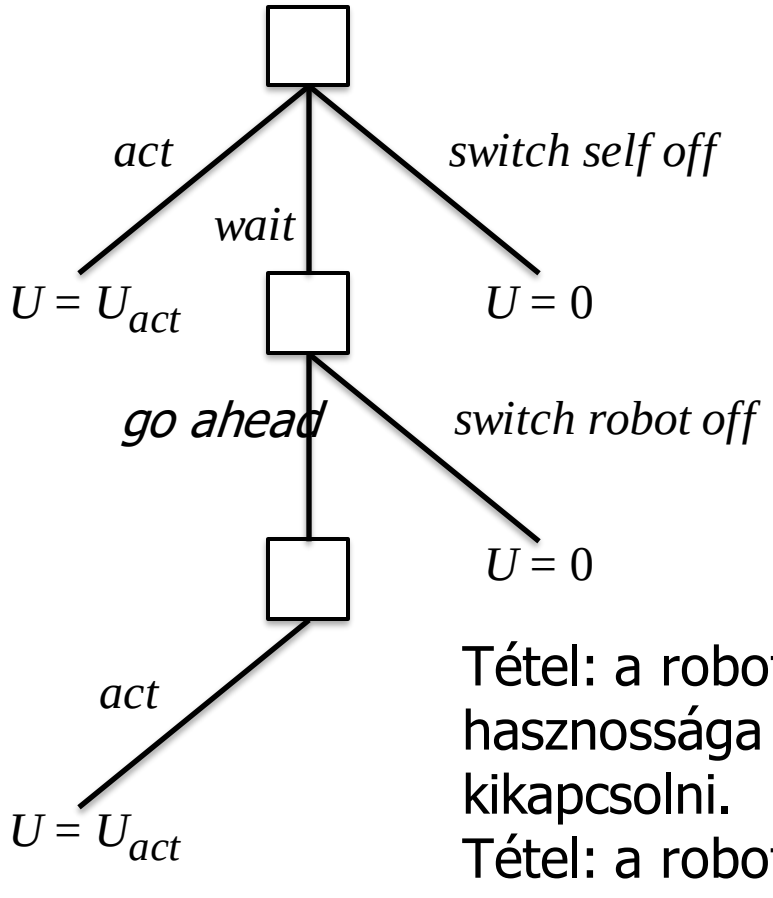
Robot



Ember(iség)
(konszenzus U?, J. Harsányi!)



Robot



Tétel: a robotnak pozitív várható hasznossága van arra, hogy hagyja magát kikapcsolni.
Tétel: a robot bizonyíthatóan jóravaló

Lehetséges problémák: inkonzisztens/zajos/rosszindulatú ember, korlátos MI reprezentáció

További információ értéke

Vezessük be a következő jelölést:

- Evidenciák (biztos ismeretek): E
- Akciók: a , legjobb akció E esetében: α
- Kimenetelek: S_i
- Potenciális további evidencia: E_j, lehetséges értékei e_{jk}

$$EU(\alpha|E) = \max_a \sum_i U(S_i)\ P(S_i|E,a)$$

- Legjobb akció további e_{jk} mellett: $\alpha_{e_{jk}}$

$$EU(\alpha_{e_{jk}}|E, E_j = e_{jk}) = \max_a \sum_i U(S_i)\ P(S_i|E,a,E_j = e_{jk})$$

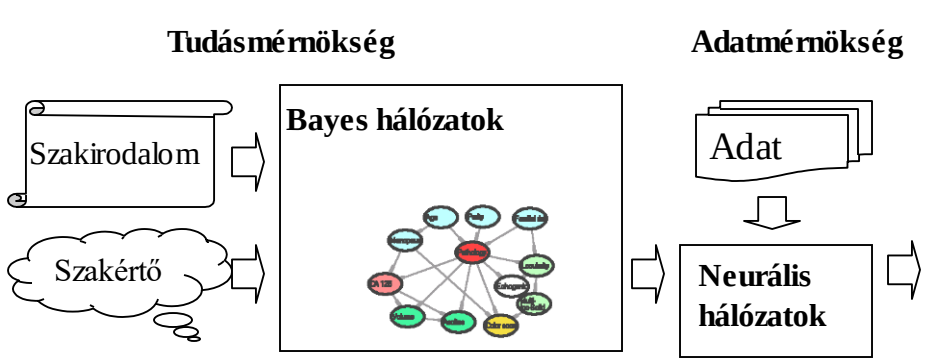
- További evidencia értéke: Várható hasznosságok különbözősége

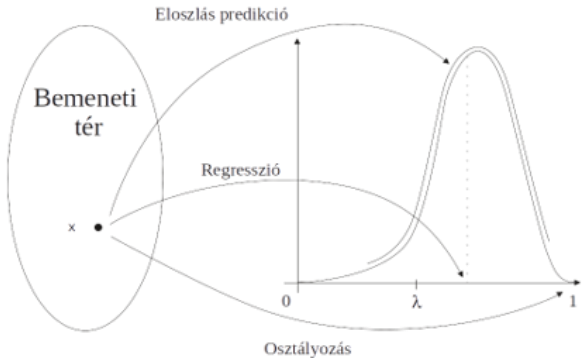
$$VPI_E(E_j) = \left(\sum_k P(E_j = e_{jk}|E) EU(\alpha_{e_{jk}}|E, E_j = e_{jk})\right) - EU(\alpha|E)$$

Russell, S. J., and Peter Norvig. "Artificial Intelligence: A Modern Approach, 4th, Global ed." (2022).

Osztályozás bizonytalansággal

Ha a loss bizonytalan és elutasítás is lehetséges, teljes bayesi predikció segíthet.





Antal, P., Fannes, G., Timmerman, D., Moreau, Y. and De Moor, B., 2003. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artificial intelligence in medicine*, *29*(1-2), pp.39-60.

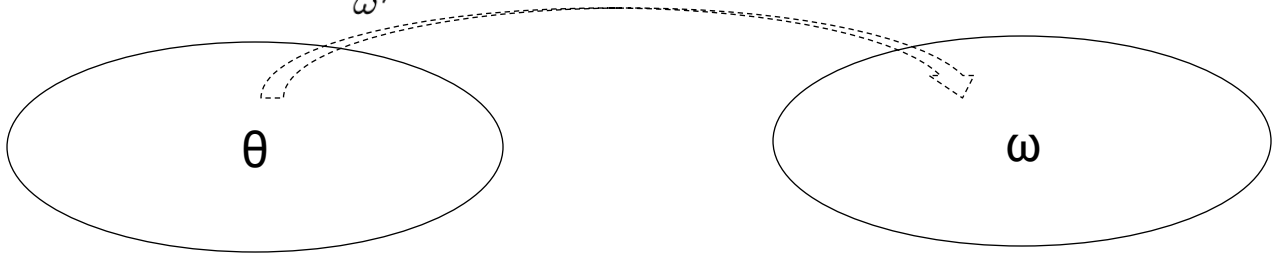
Emberközpontú MI: értelmezhetőség

Modellek leképzése és transzformációja

Célok

- Modellek egyszerűsítése/felskálázása
- A priori információk bevitele
- Regularizáció
- Számítási komplexitás csökkentése (idő, energia)
- Értelmezhetőség (verifikáció, validáció, auditálás)

$$\mathcal{T}_{\mathrm{KL}}(\theta) = \arg\min_{\omega'} \mathrm{E}_{p(X|\theta)}[\mathrm{KL}(p(Y|X,\omega')\|p(Y|X,\theta))] + c(\omega)$$



Antal, P., Fannes, G., Verrelst, H., De Moor, B. and Vandewalle, J., 2000. **Incorporation of prior knowledge in black-box models: comparison of transformation methods from Bayesian network to multilayer perceptrons**. In *Proc. of the Workshop on Fusion of Domain Knowledge with Data for Decision Support, The Sixteenth Conference on Uncertainty in Artificial Intelligence* (pp. 7-12).

Bayesian mapping between black-box and white-box models

1.BN parameter generation by independent Dirichlet sampling,

2.Data generation

- •Forward sampling =>i.i.d.
- •Computing the output class probabilities

conditioned only on the sampled input values (PPTC)

=> less sample, higher accuracy

3.MLP training

- •Scaled Conjugate Gradient with weight decay =>fast, automatic optimization method for complex models

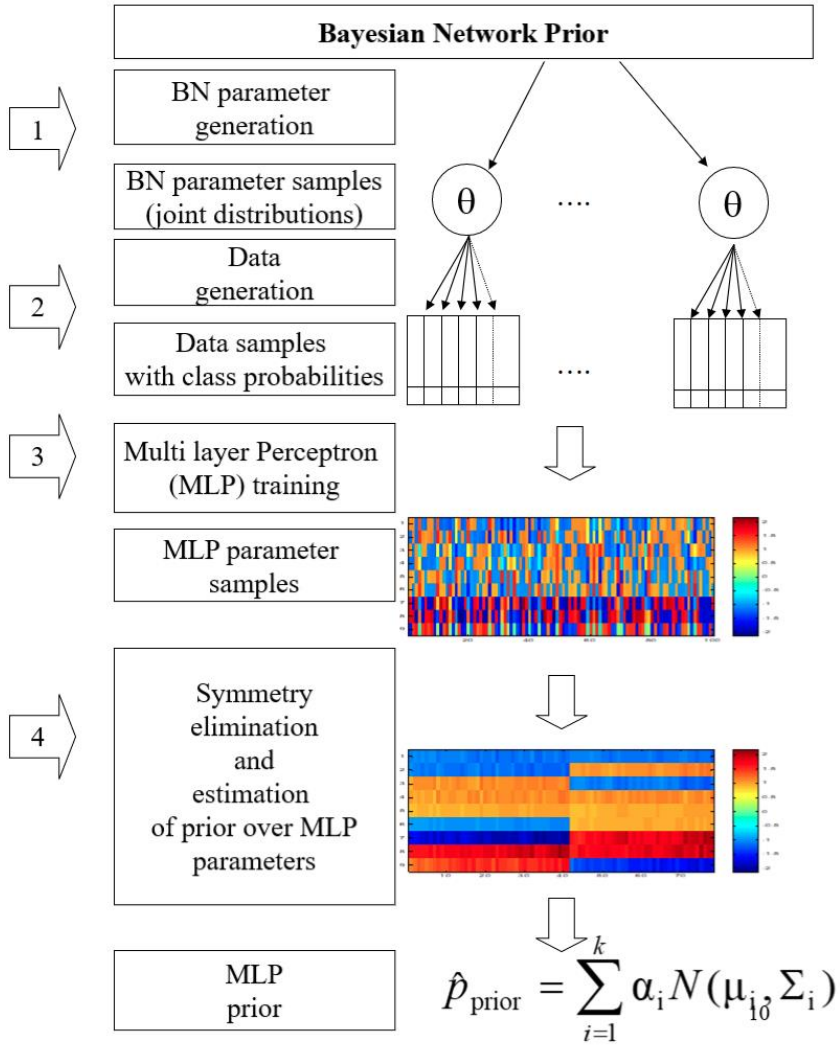
=>penalty standardize the behavior of optimization producing smoother and more concentrated sampling for density estimation

- •Stucking detection based on the distribution of training error

4.Prior density estimation over MLP parameter space

- •Estimation by symmetry elimination and clustering

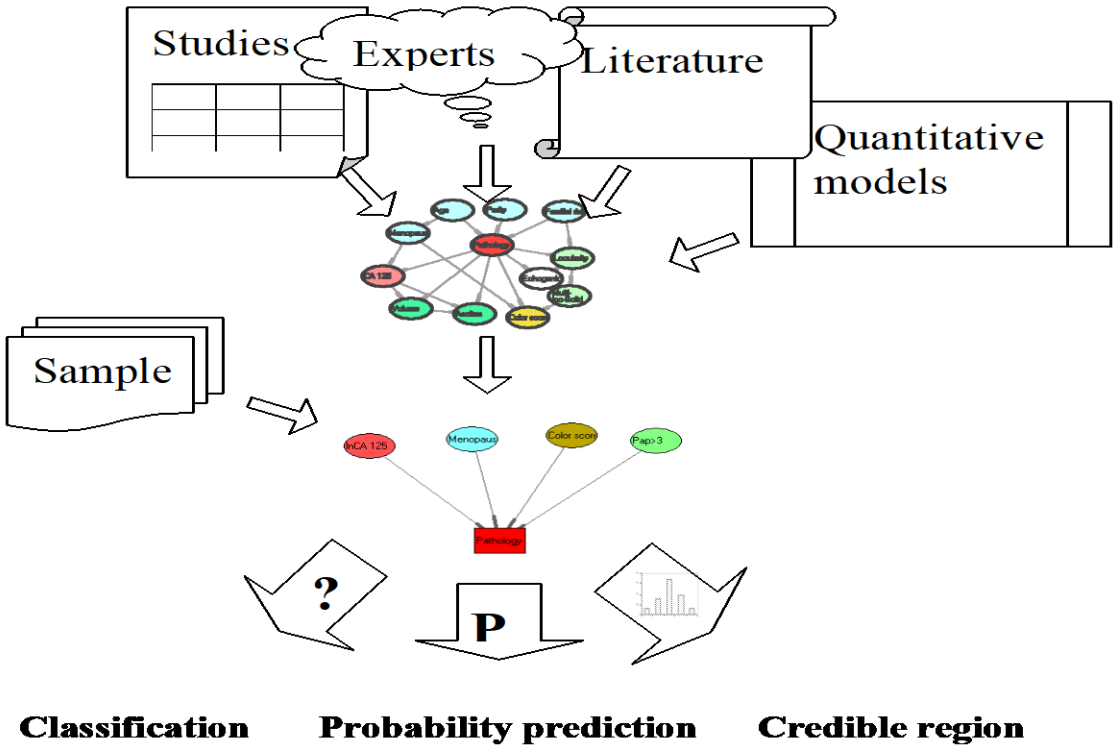
5.Bayesian inference using derived informative prior and data



Antal, P., Fannes, G., Verrelst, H., De Moor, B. and Vandewalle, J., 2000. **Incorporation of prior knowledge in black-box models: comparison of transformation methods from Bayesian network to multilayer perceptrons**. In *Proc. of the Workshop on Fusion of Domain Knowledge with Data for Decision Support, The Sixteenth Conference on Uncertainty in Artificial Intelligence* (pp. 7-12).

Emberközpontú MI: magyarázatgenerálás

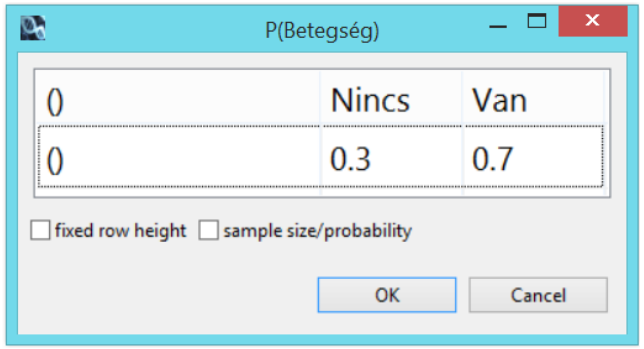
AI for evidence-based medicine

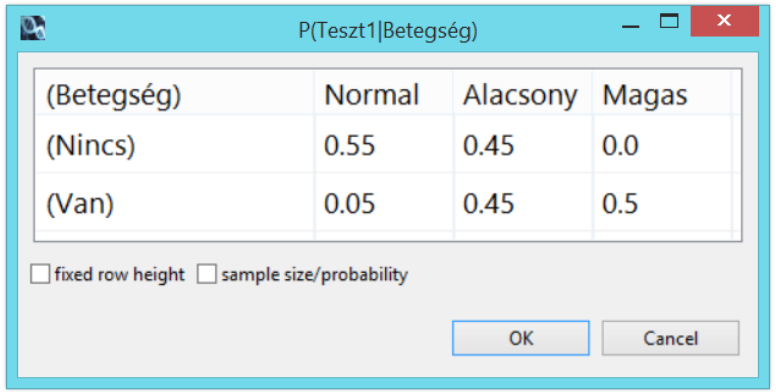


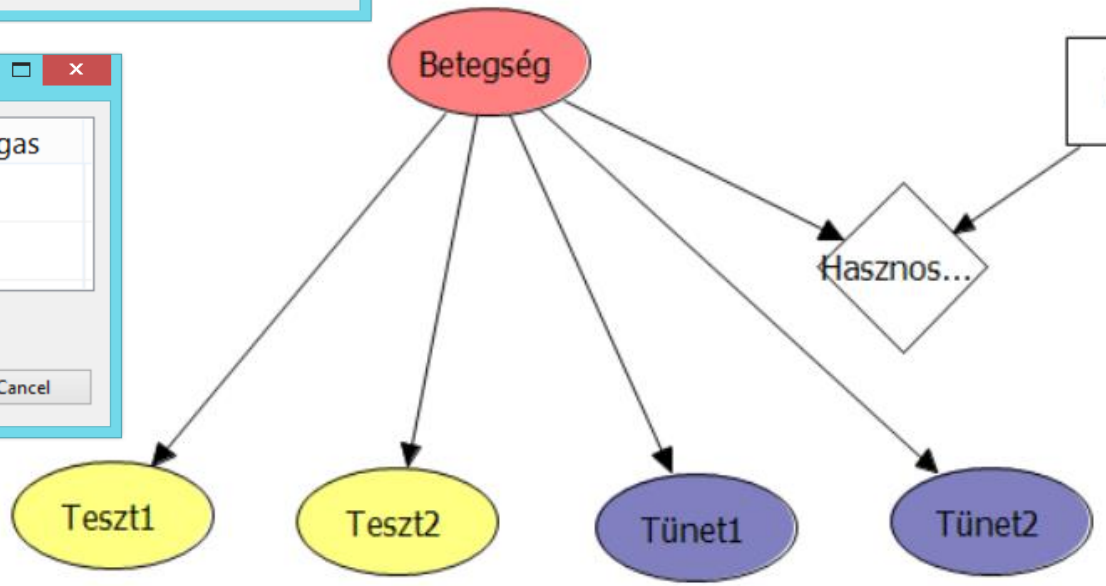
International Ovarian Tumor Analysis (IOTA, Dirk Timmerman)

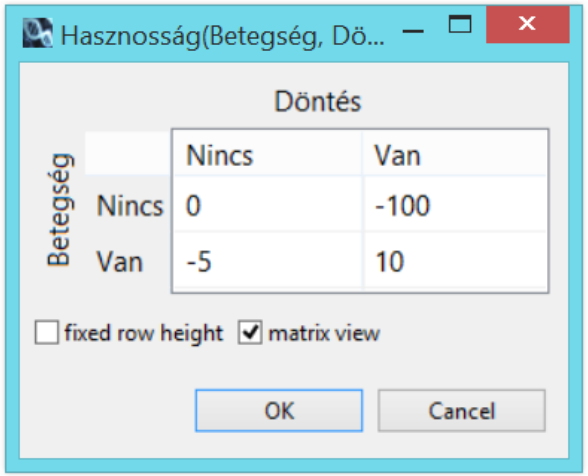
P. Antal, G. Fannes, D. Timmerman, Y. Moreau, B. De Moor: Bayesian Applications of Belief Networks and Multilayer Perceptrons for Ovarian Tumor Classification with Rejection, *Artificial Intelligence in Medicine*, vol. 29, pp 39-60, 2003

Döntési (naív) hálózat ("idiot Bayes network")









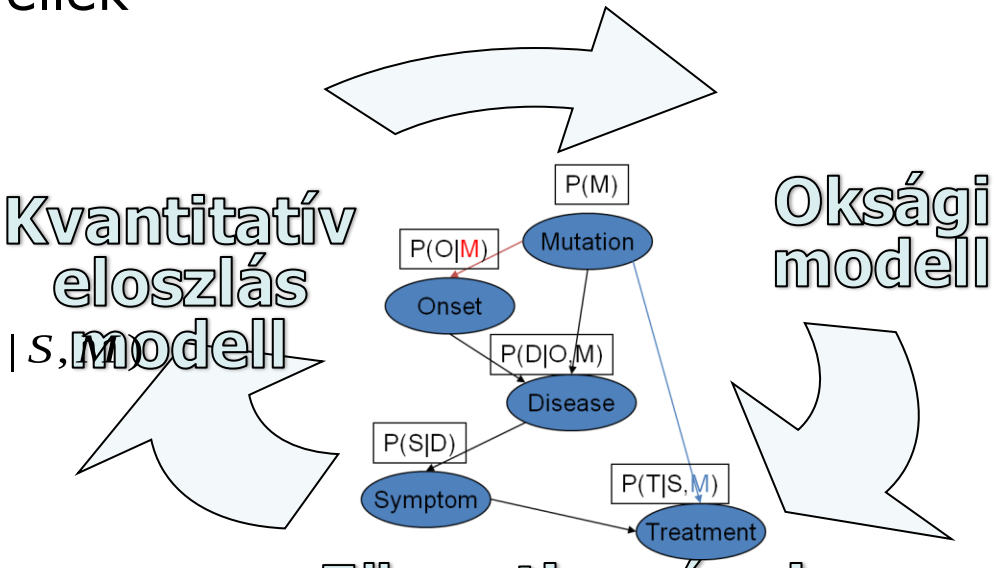
Bayes-hálózatok

Valószínűségi gráfos modell

- csomópont – véletlen változók
- él – közvetlen függés/ (oksági kapcsolat)
- lokális valószínűségi modellek

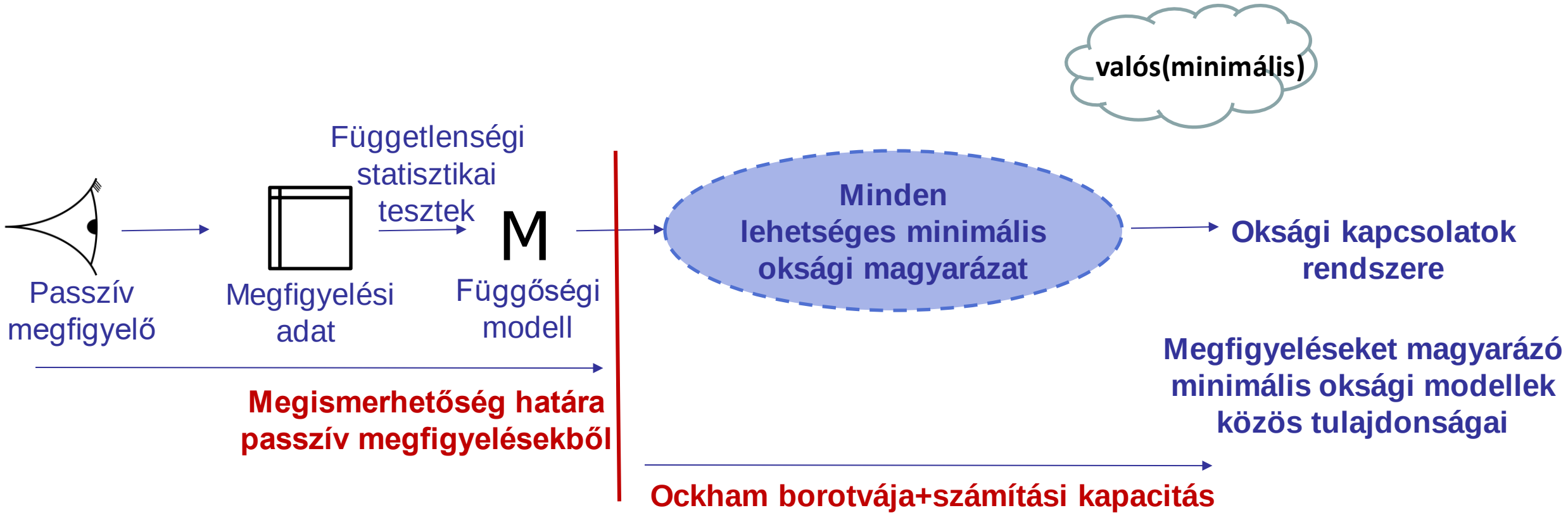
Három értelmezés

$$P(M,O,D,S,T) = P(M)P(O|M)P(D|O,M)P(S|D)P(T|S,M)$$

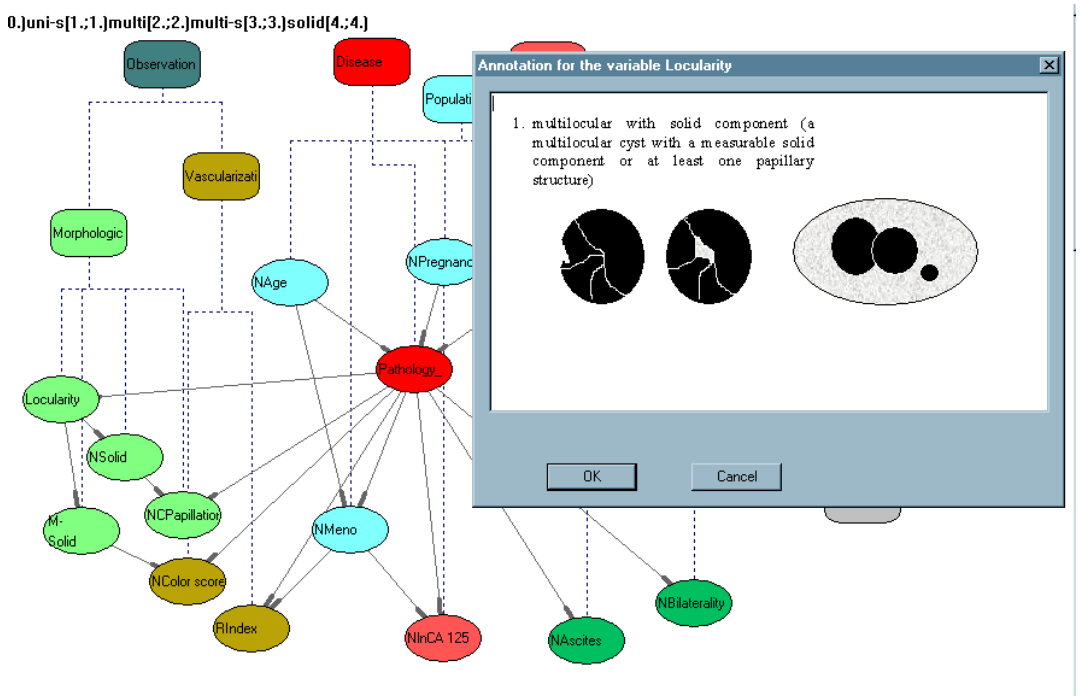


$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$

Indukció okságra

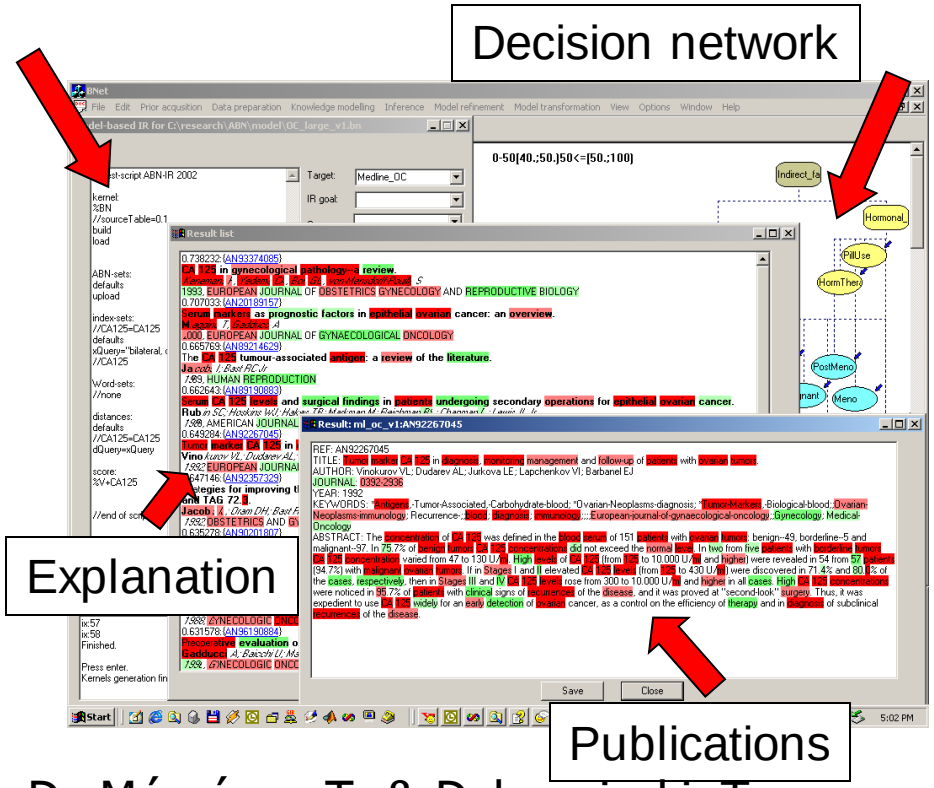


Semantically enriched evidence-based models



Antal, P., Mészáros, T., De Moor, B., & Dobrowiecki, T. (2001). Annotated Bayesian Networks: a tool to integrate textual and probabilistic medical knowledge. In *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001* (pp. 177-182). IEEE.

Most probable annotated causal explanations



Antal, P., De Moor, B., Timmerman, D., Mészáros, T., & Dobrowiecki, T. (2002). Domain knowledge based information retrieval language: an application of annotated Bayesian networks in ovarian cancer domain. In *Proceedings of 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)* (pp. 213-218). IEEE.

A döntéstámogatás típusai

Valószínűségszámítás, statisztika

Diagnosztika

Okozatiság

Beavatkozás?

Optimális cselekvés?

Mi lett volna, ha...?

Döntéselmélet

Intelligens érvelés

A döntéstámogatás típusai

- Diagnosztikai következtetés
  - P(Diagnózis|Passzív megfigyelések)
- Optimális információgyűjtés
  - P(Diagnózis|megfigyelések, új megfigyelés)
  - További információ hasznossága
- Terápiás következtetés
  - P(Kimenetel|Megfigyelés, Beavatkozás)
  - Terápia hatása
- Optimális döntés
  - Maximális hasznosságú (elérhető) beavatkozás megválasztása
- Kontrafaktuális következtetés
  - P(ElképzeltKimenetel|Megfigyelés, Beavatkozás,Kimenetel, ElképzeltBeavatkozás)

A megmagyarázható MI szintjei

Valószínűségszámítás
Statisztika
Függetlenségek rendszere
Következtetés, asszociatív predikció
Legvalószínűbb magyarázat

Oksági kutatás

Oksági felfedezés
Oksági következtetés
Kísérlettervezés

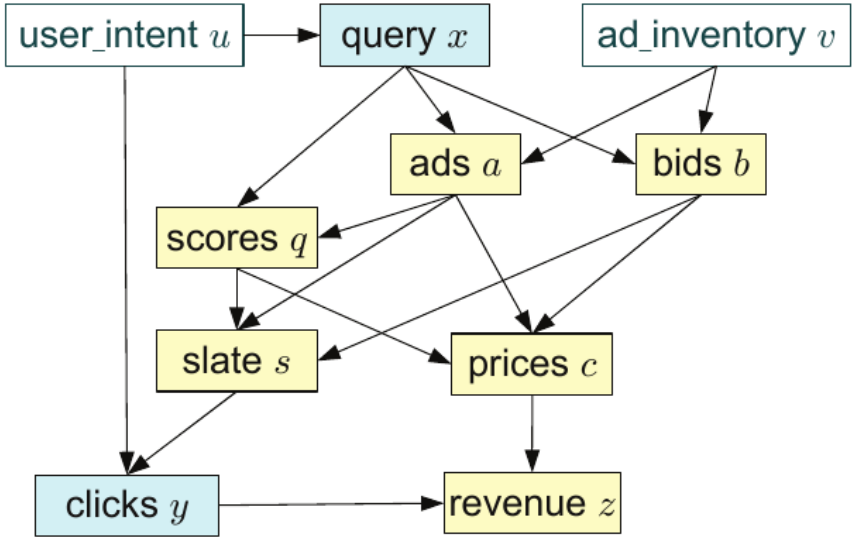
Döntéselmélet

Döntési hálózatok
Optimális döntés
Szekvenciális döntések
Megerősítéses tanulás

Intelligens érvelés

Kontrafaktuális érvelés

Online hirdetési példa



Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D.X., Chickering, D.M., Portugaly, E., Ray, D., Simard, P. and Snelson, E., 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, *14*(11).

Kontrafaktuális igazságosság

Igazságos, ha védett információ explicit módon nem használt a döntésben.

Definition 1 (Fairness Through Unawareness (FTU)). *An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process.*

Igazságos, ha védett információ nem befolyásolja a döntést.

Definition 2 (Individual Fairness (IF)). *An algorithm is fair if it gives similar predictions to similar individuals. Formally, given a metric $d(\cdot, \cdot)$, if individuals i and j are similar under this metric (i.e., $d(i, j)$ is small) then their predictions should be similar: $\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)})$.*

Igazságos, ha védett információ elképzelt világokban sem befolyásolja a döntést.

Definition 5 (Counterfactual fairness). *Predictor $\hat{Y}$ is* **counterfactually fair** *if under any context $X = x$ and $A = a$,*

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \quad (1)$$

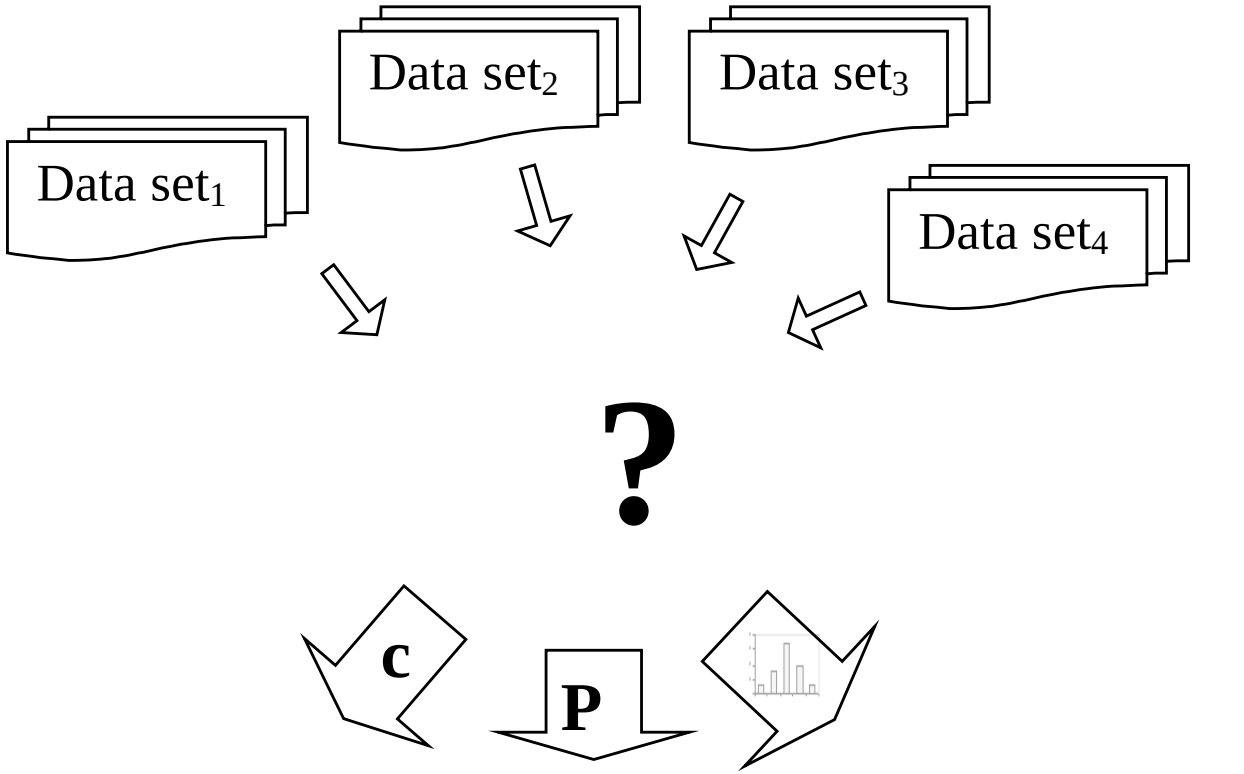
for all y and for any value a' attainable by A.

Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. *Advances in neural information processing systems*, *30*.

Baer, B.R., Gilbert, D.E. and Wells, M.T., 2020. Fairness criteria through the lens of directed acyclic graphs. In The Oxford Handbook of Ethics of AI.

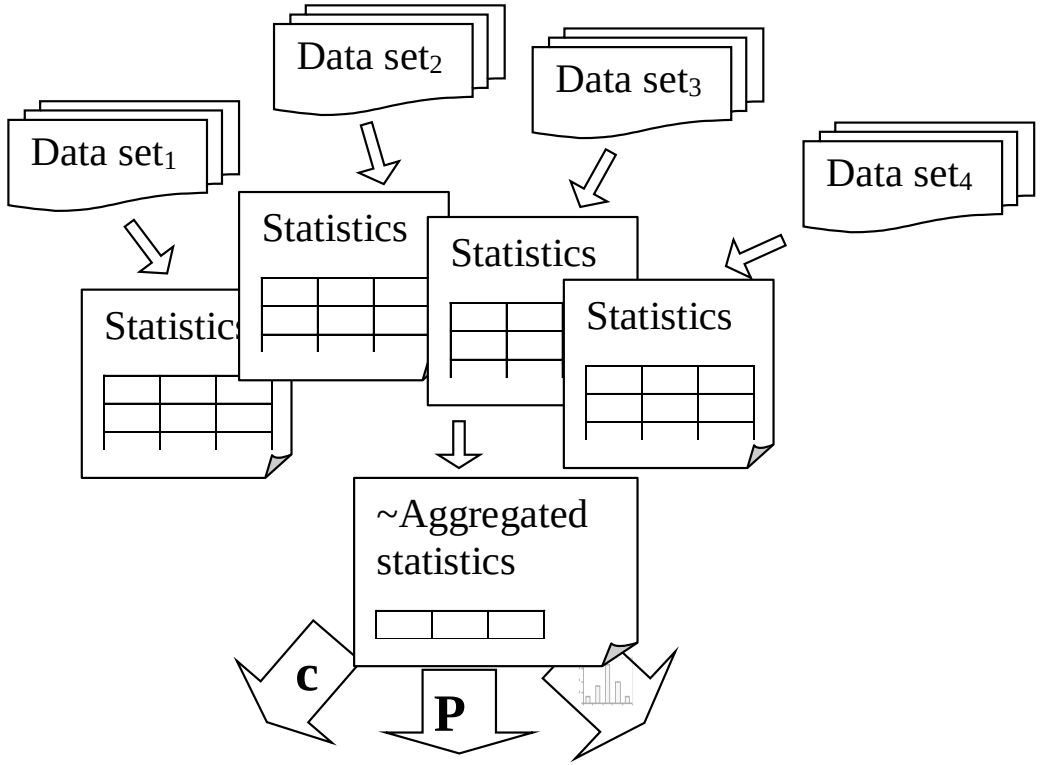
Federált MI

Challenge: learning from multiple data sets



Classifications **Probability predictions** **Credible regions**

Meta-analysis



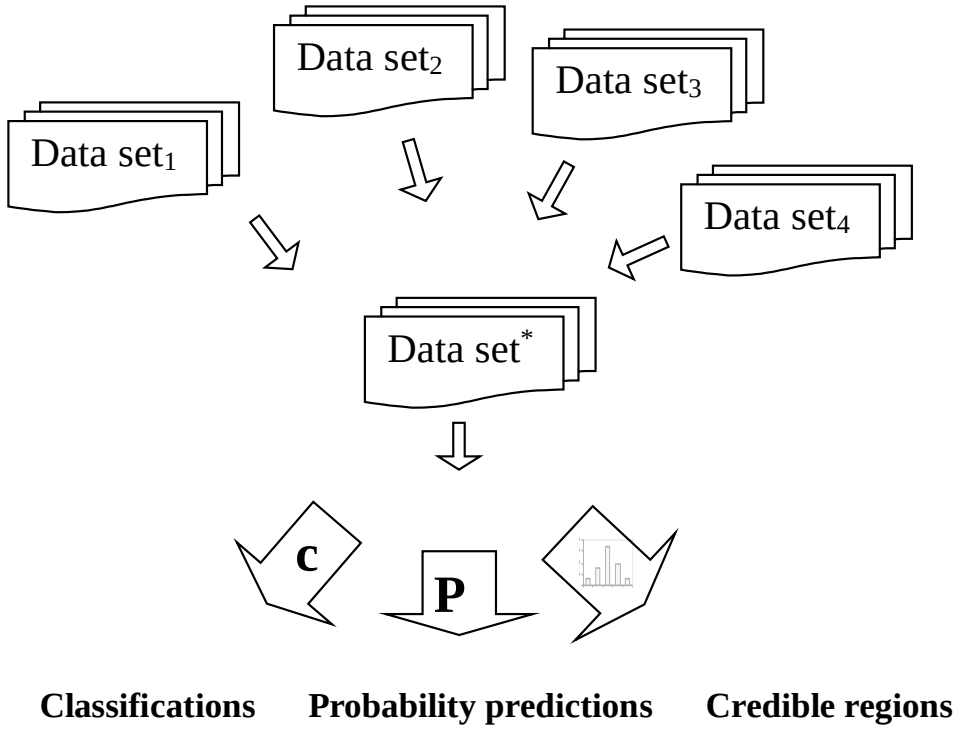
Classifications **Probability predictions** **Credible regions**

Challenges:
- Statistics
- Privacy

Evidence-based medicine: The Cochrane (Collaboration)

https://www.cochrane.org/

Learning from pooled data



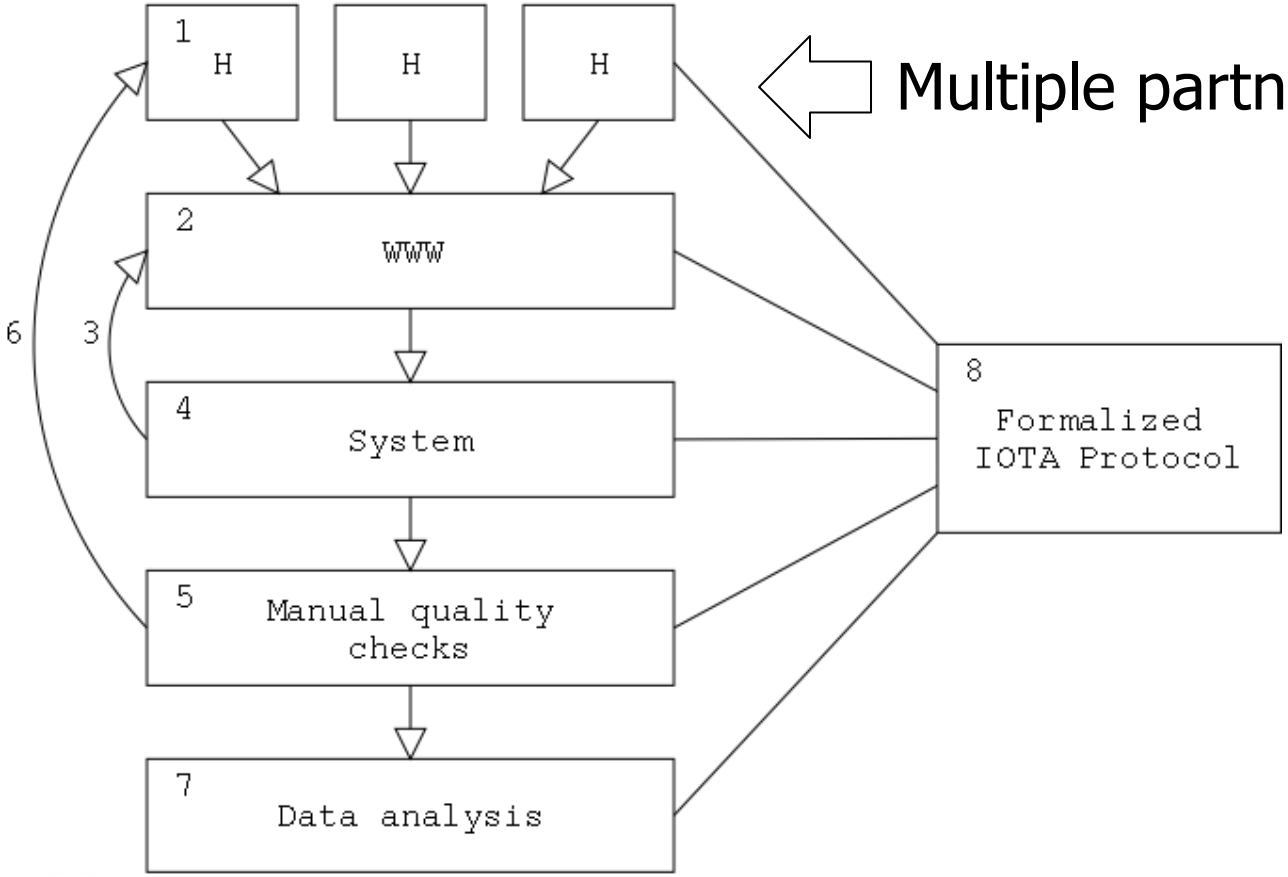
Challenges:

- **Data harmonization**
- Privacy
- Communication cost
- Centralized storage cost
- Centralized computation cost

IOTA: a multi-centric study

Non-invasive diagnostics of ovarian cancer, 1998

International Ovarian Tumor Analysis (IOTA, http://www.iotagroup.org/)

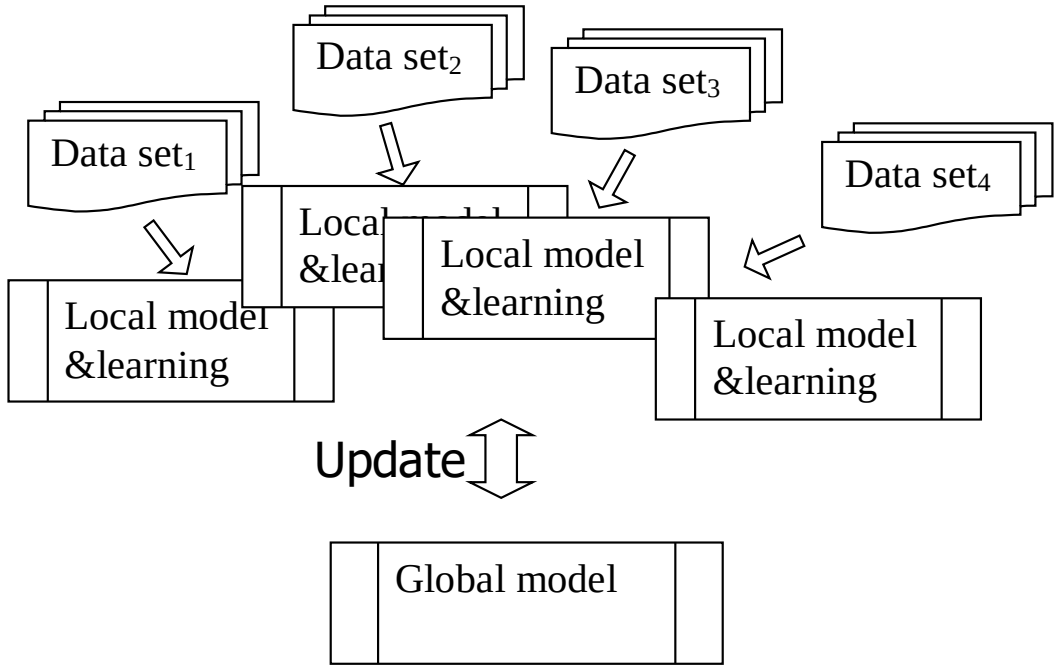


Aerts, S., Antal, P., Timmerman, D., De Moor, B. and Moreau, Y., 2002, June. Web-based data collection for uterine adnexal tumors: a case study. In *Proceedings of 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)* (pp. 282-287). IEEE.

Federated learning from multiple local data sets

Separation of data and model/learning

1. **Data is harmonized**
2. Stays at the institutes/**individuals**
3. Model updates are communicated
4. Using privacy-preserving techniques



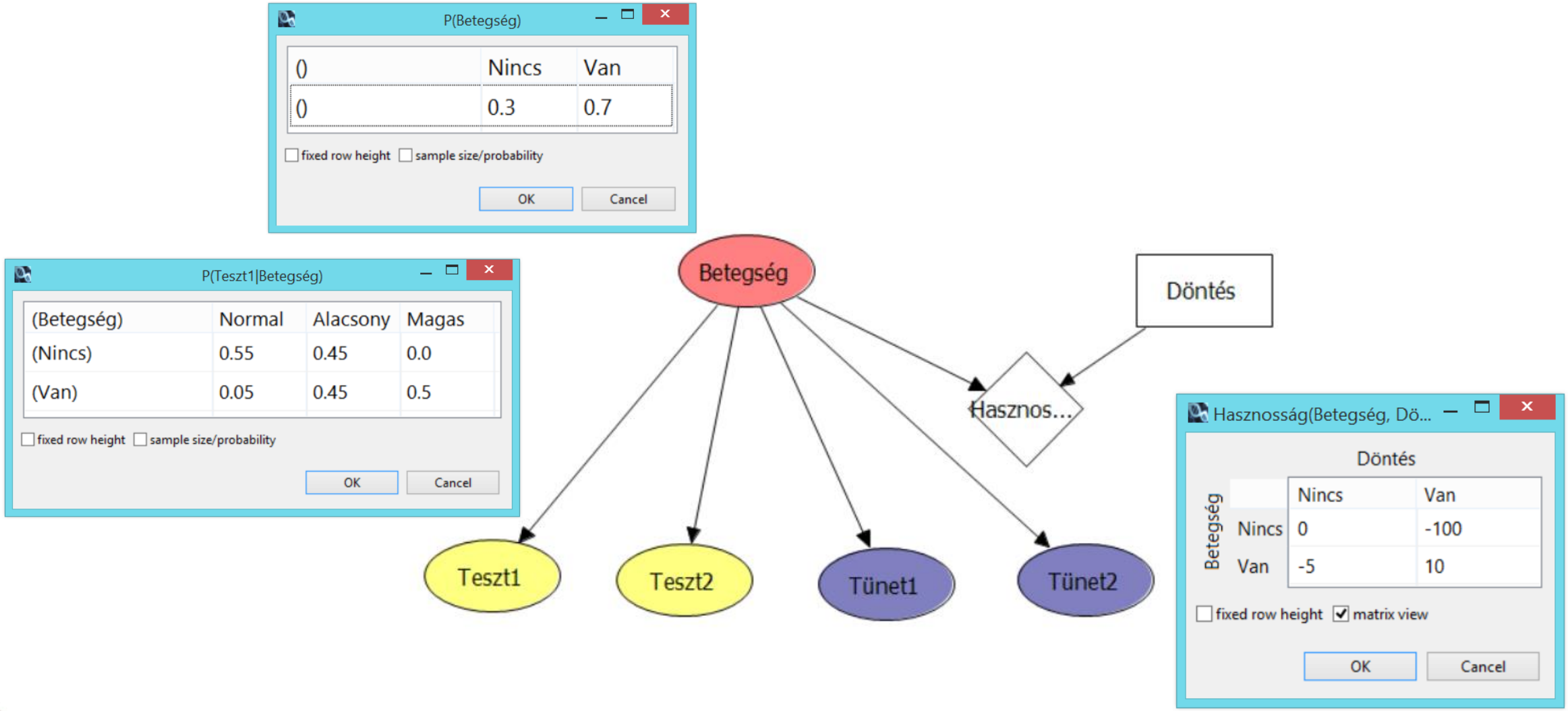
Konečný, J., McMahan, B. and Ramage, D., 2015. Federated optimization: Distributed optimization beyond the datacenter.
Konečný, J., McMahan, H.B., Ramage, D. and Richtárik, P., 2016. Federated optimization: Distributed machine learning for on-device intelligence.
McMahan, H.B., Moore, E., Ramage, D. and Hampson, S., 2016. Communication-efficient learning of deep networks from decentralized data.

Federated learning (FL): a definition

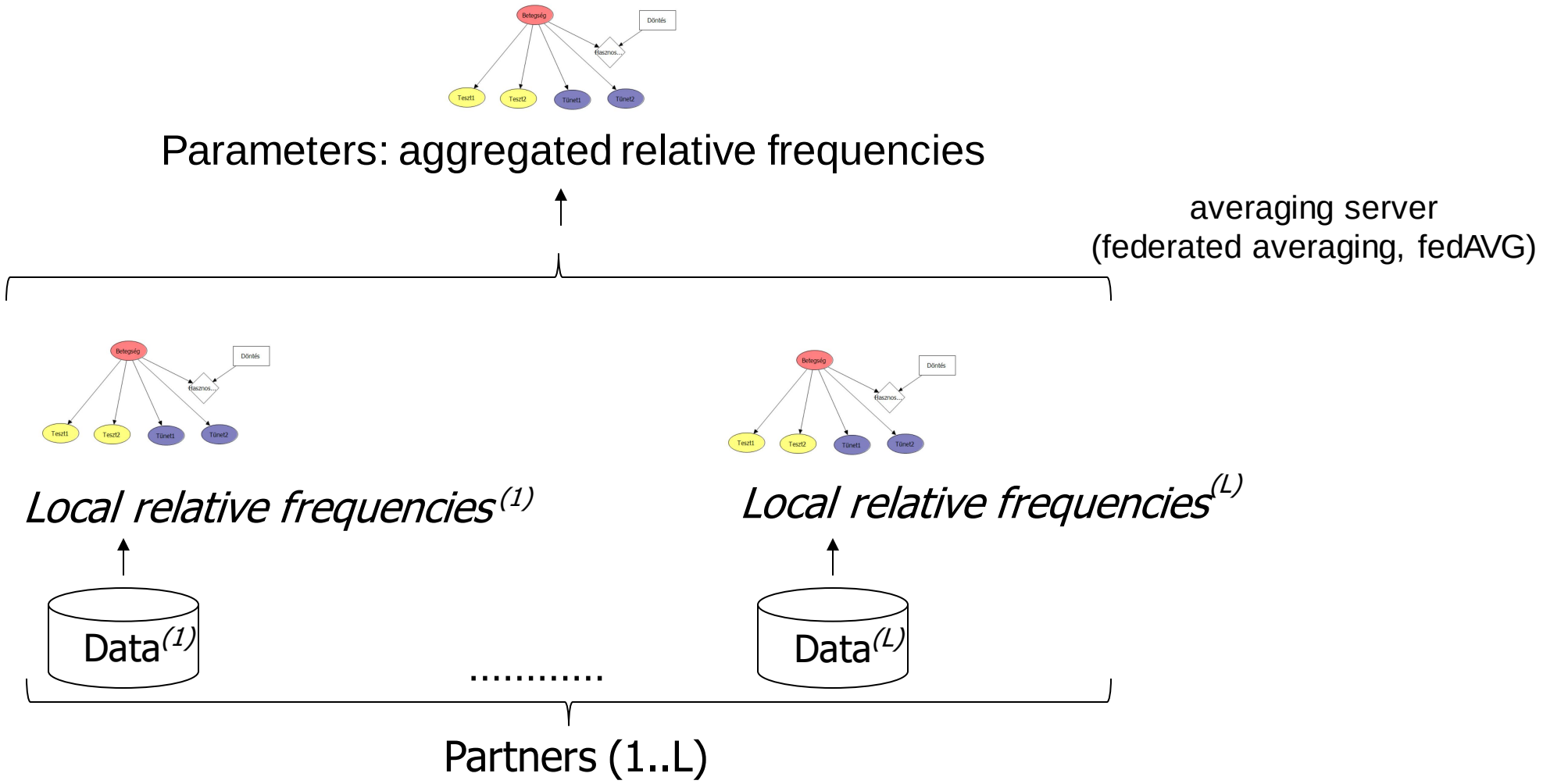
Federated learning is a machine learning setting where **multiple entities (clients) collaborate** in solving a machine learning problem, **under the coordination of a central server** or service provider. Each **client's raw data is stored locally** and not exchanged or transferred; instead, **focused updates intended for immediate aggregation are used** to achieve the learning objective.

Kairouz, Peter, et al. "Advances and open problems in federated learning." arXiv preprint arXiv:1912.04977 (2019).

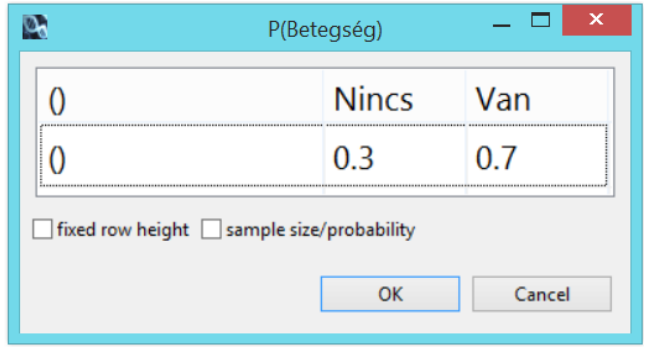
Döntési (naív) hálózat

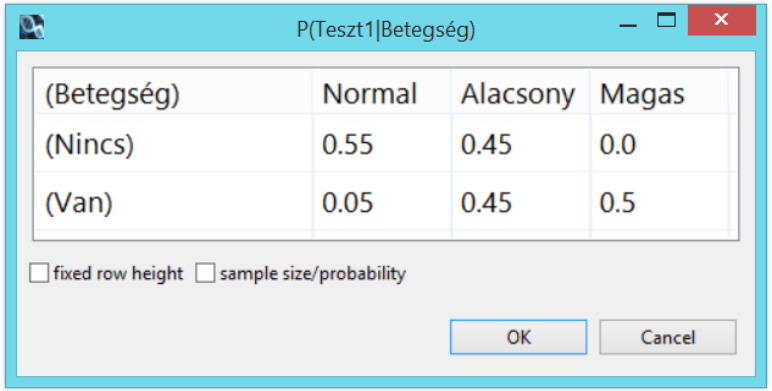


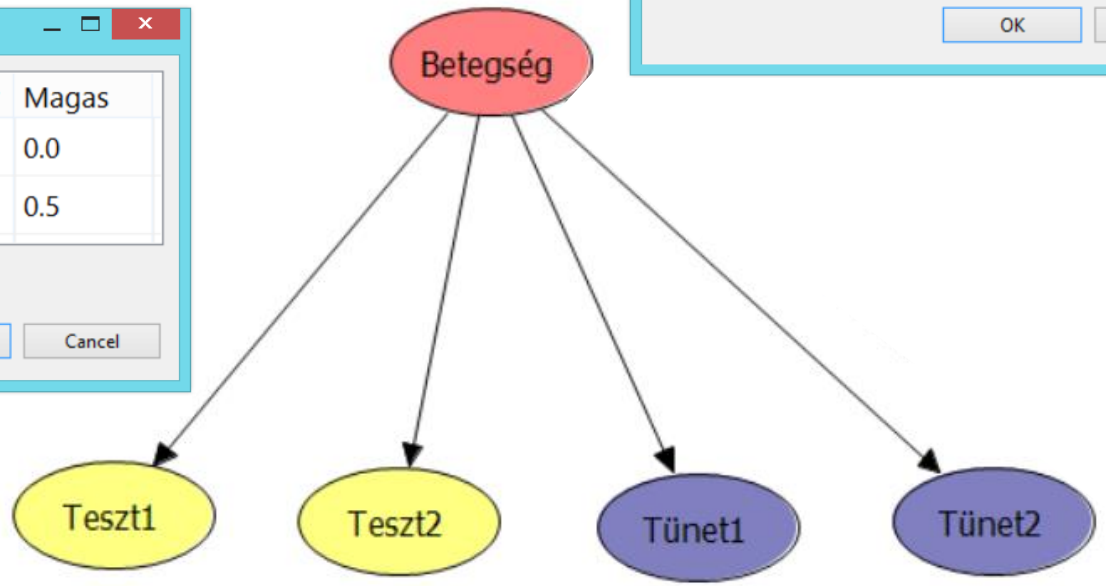
Federated learning: exact parameter averaging



Naiv Bayes hálózat



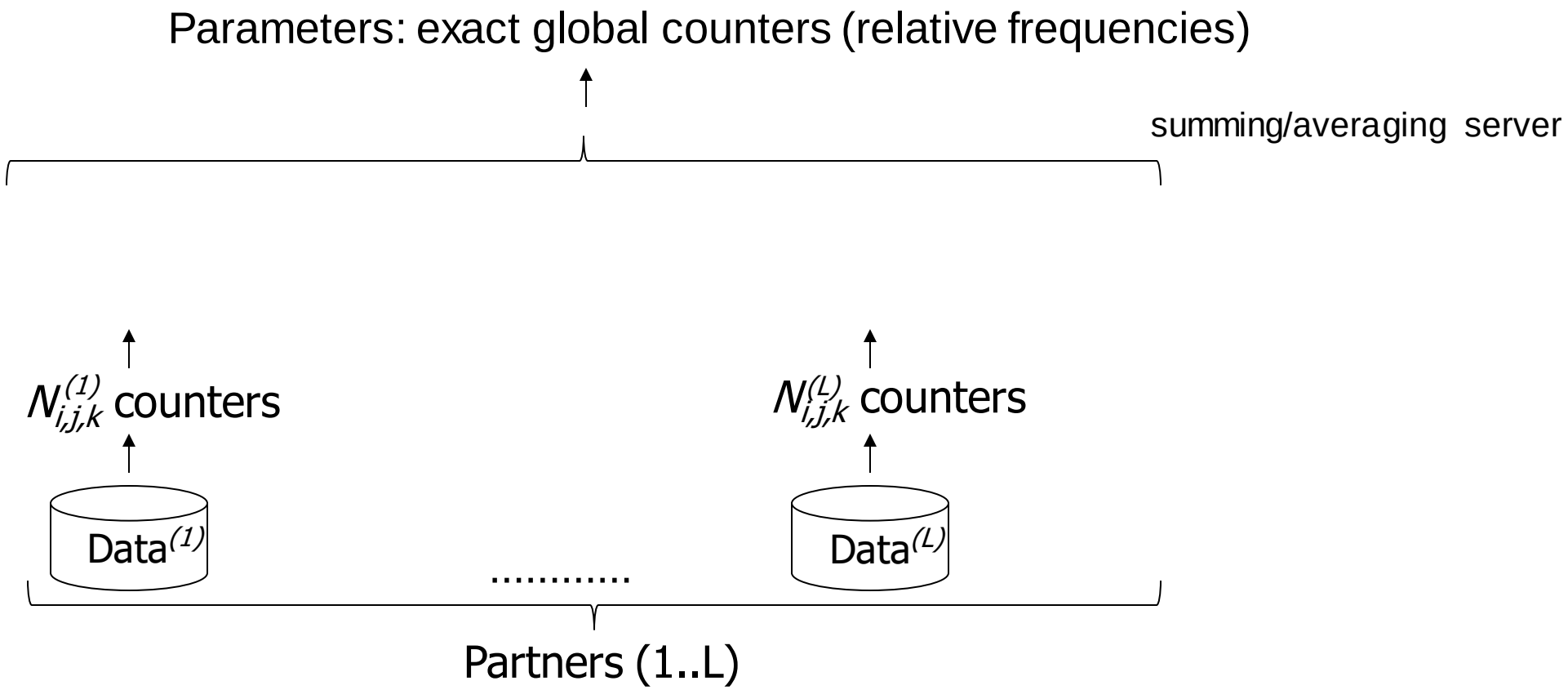




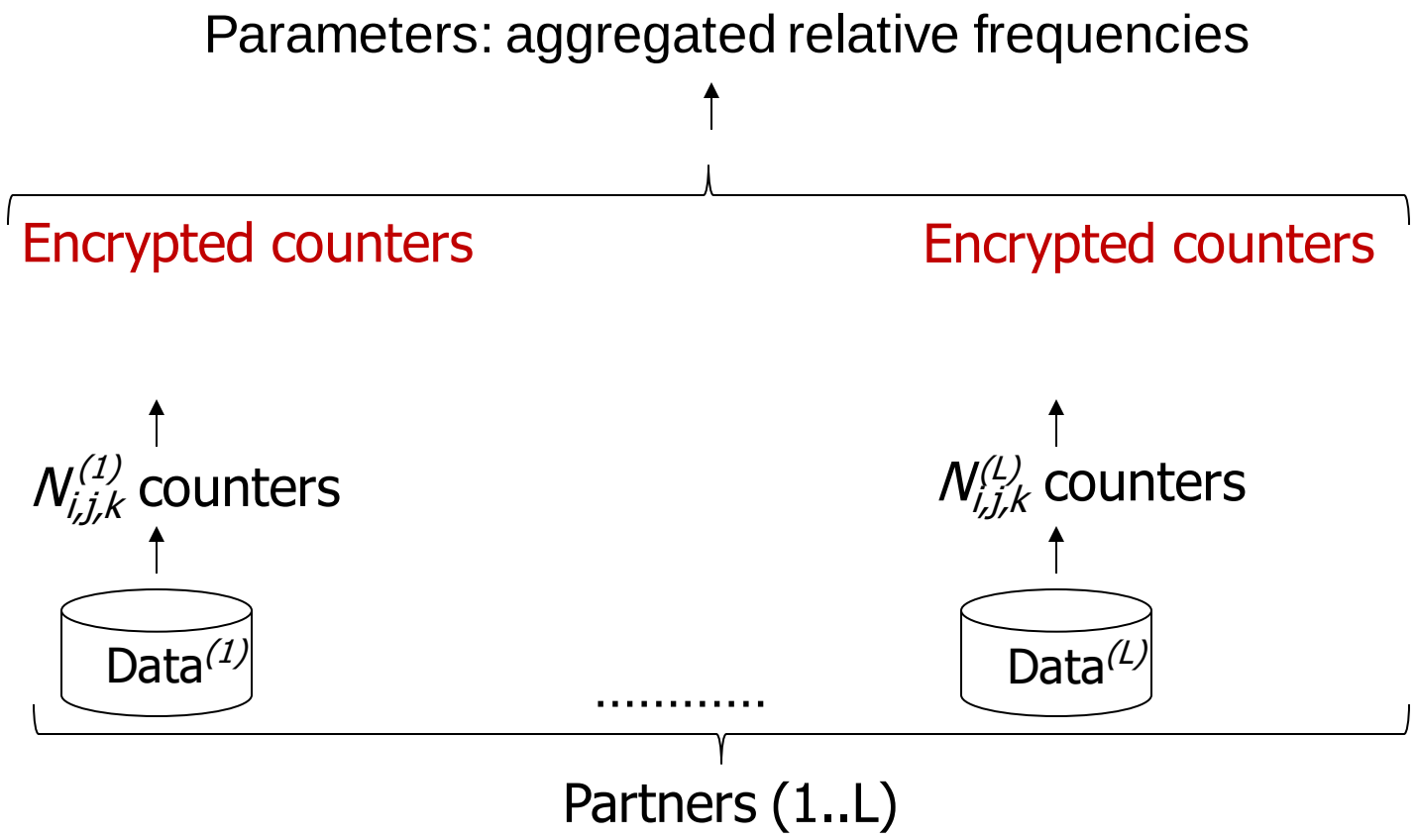
P(Betegség=van)~N(Betegség=van)/N
P(T1=Normal|B=van)~N(T1=Normal,B=van)/N(B=van)

Multinomiális eloszláshoz a számlálók *elégséges statisztikák*.

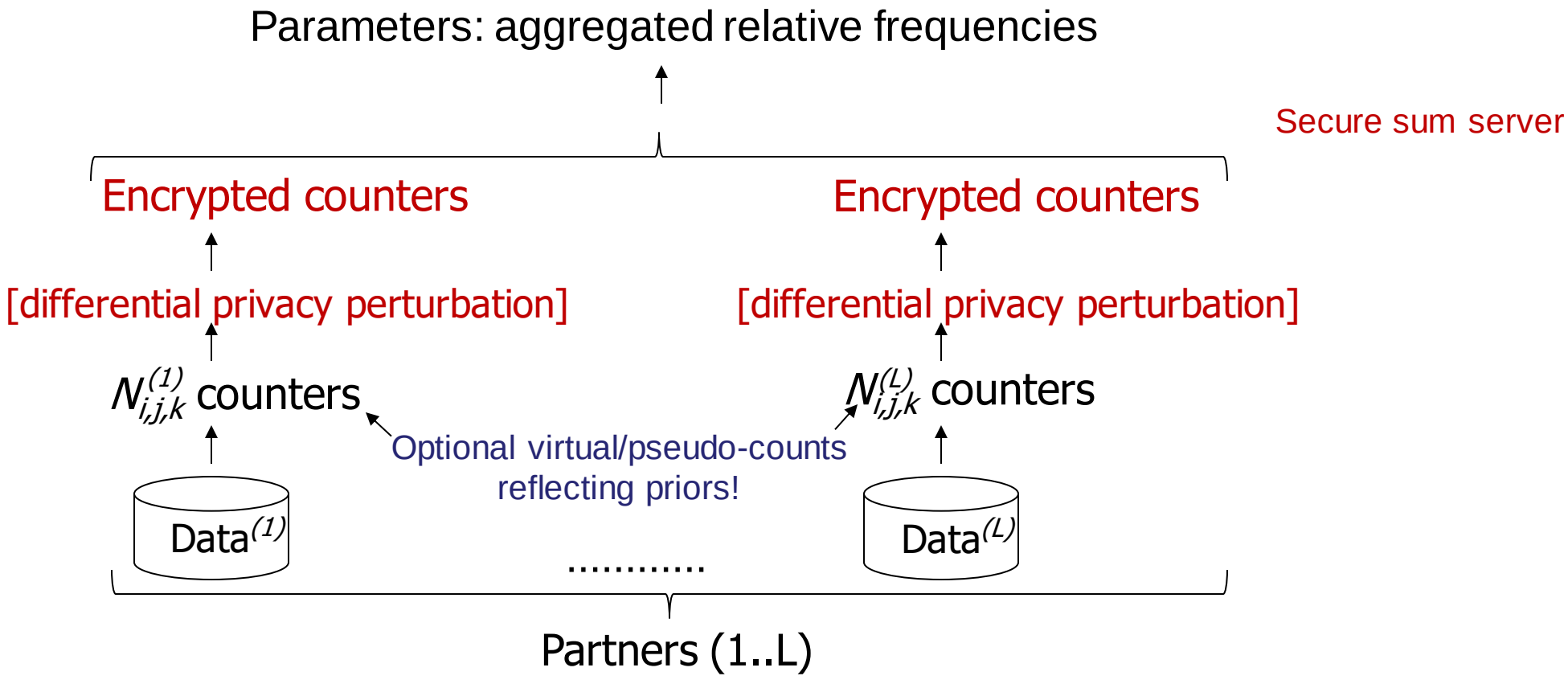
Federated learning: exact parameter learning



Privacy-preserving federated learning



Privacy-preserving federated learning



Advantages and challenges

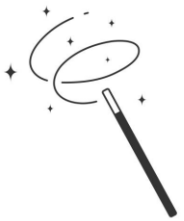
- Data stays at the owner
  - privacy
  - maintainance
- Computation stays at the owner
  - limited overhead: server+communication
- Method stays at the owner
  - settings ("hyperparametrization")
- The global model is based on
  - "all" data,
  - "all" compute power,
  - "all" local fine tunings

  at the partners.

- Data harmonization
  - throughout the lifecycle(!)
- Communication cost
  - limiting factor for compute power
- Heterogenous (non-iid) data sets:
  - partner specific vs. global model
- Privacy of local data sets
  - selection of non-optimal/useless statistical models
- Security and valorization of the global model
  - evaluation of contribution from the partners

Összefoglalás

- Az MI veszélyes (és egyre veszélyesebbé váló) eszköz, de
- lehetséges **bizonyíthatóan jóravaló MI**
  - emberhez igazodás preferenciák rögzítése nélkül
- lehetséges egzotikus **modellek értelmezése**
  - Bayesi leképzések révén
- lehetséges **intelligens magyarázat**
  - akár megfigyelési adatból tanult oksági modellekkel
- lehetséges **autonómiát megőrző együttes tanulás**
  - akár érzékeny adatok és megbízhatatlan környezet esetén



azaz az EMMI lehetséges ;-), csupán "Okosabban kéne élni"