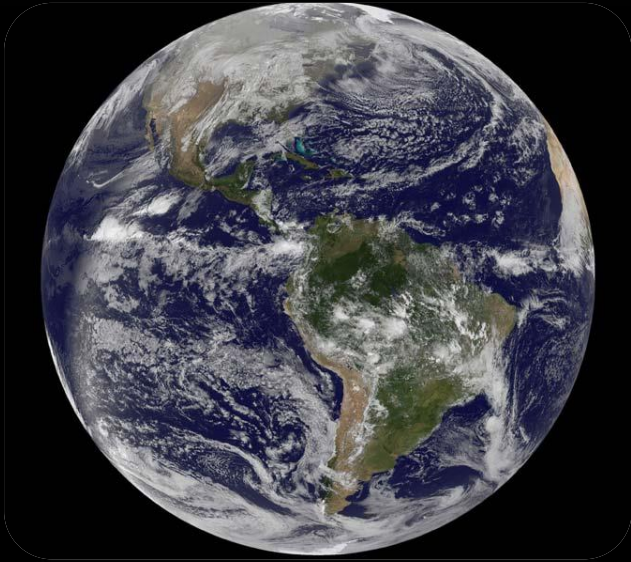


Systems Immunology & Immune Oncology

A Data-Centric View

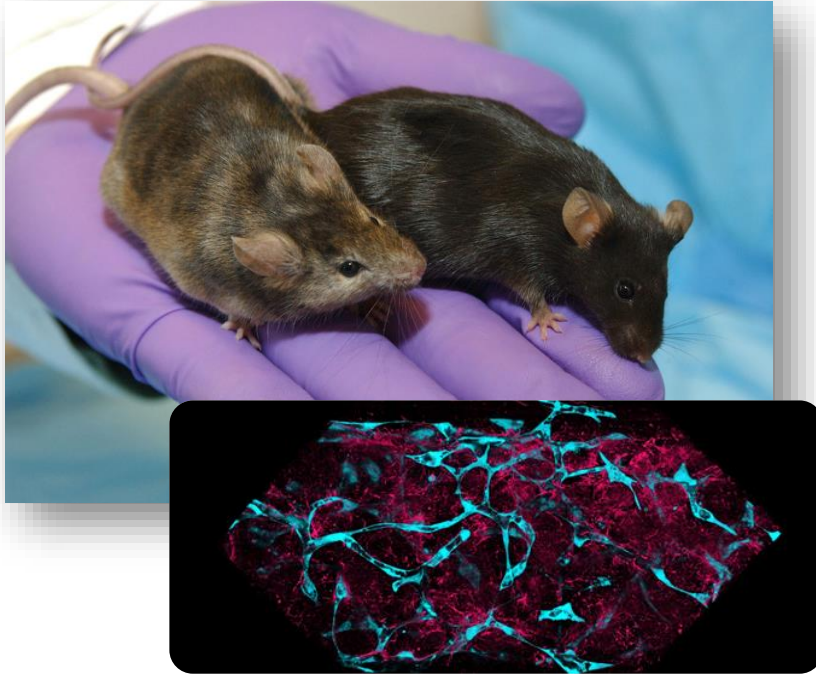


Magnus Fontes

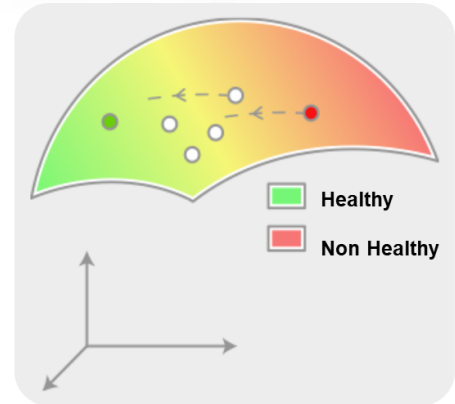
Disclosures & Affiliations

- General Manager of Institut Roche, France <https://institut.roche.com/>
- Adjunct professor of mathematics, Lund University, Sweden
<https://portal.research.lu.se/en/persons/magnus-fontes>
- Co-founder of the bioinformatics software company Qlucore
www.glucore.com
- Some images and diagrams have been created partly using Gemini or ChatGPT.

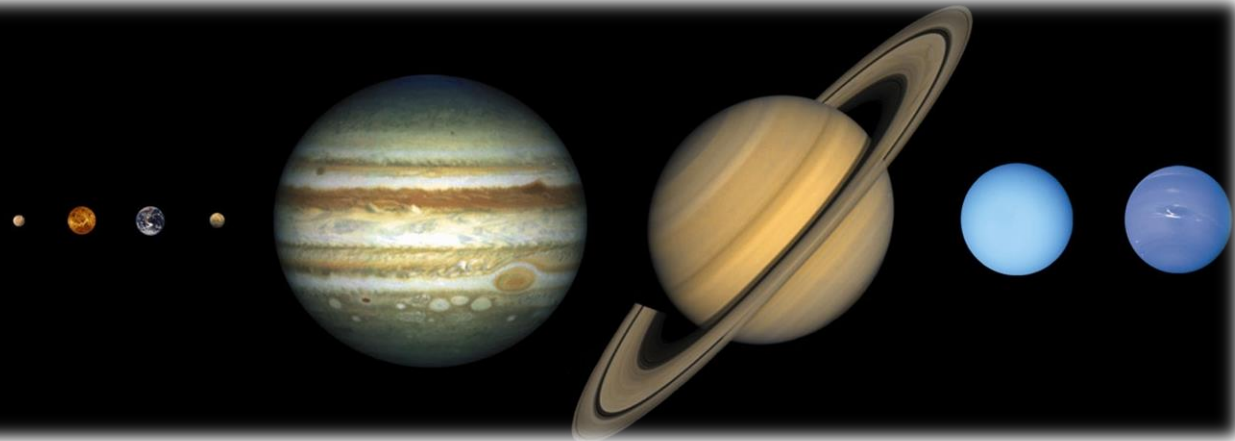
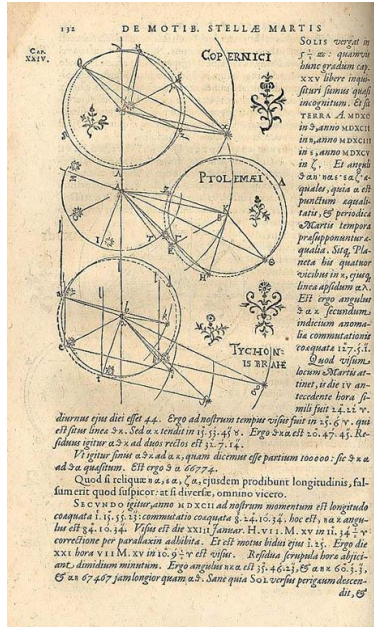
Multidisciplinary *models* to try to *understand* biology, But, what is a model?



$$\frac{dx}{dt} = \alpha x - \beta xy,$$
$$\frac{dy}{dt} = \delta xy - \gamma y,$$

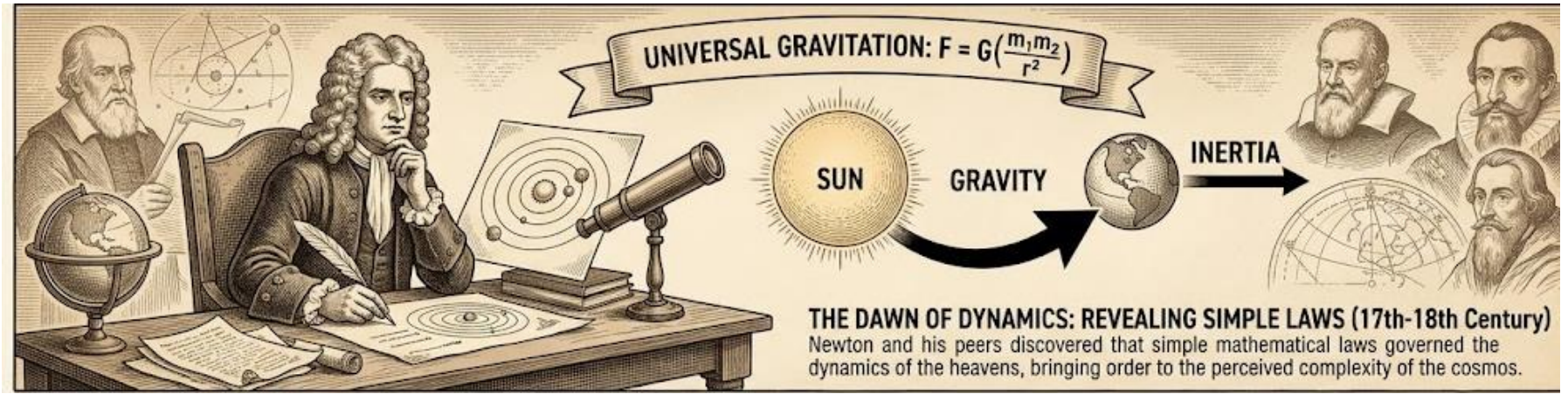


Modeling a famous System



Astronomia Nova,
Johannes Kepler, 1609

https://solarsystem.nasa.gov/multimedia/display.cfm?IM_ID=178



Are there Kepler or Newton style laws in Biology?

Mechanical dynamics are “simple” in the sense that a few characteristics and coupled measurements predict the future

Example: A Spring System:

Establish mass, stiffness and damping, measure displacement and plug into compact physical law equations → predict motion over time.



wall / reference state = position + velocity

A spring–mass system reduces motion to a few physical parameters.

Compact model

force balance

$$m\ddot{x} + c\dot{x} + kx = 0$$

state-space dynamics

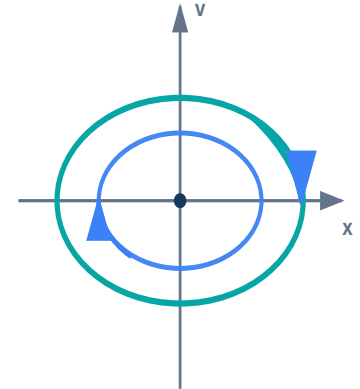
$$\dot{x} = v, \quad \dot{v} = -\frac{k}{m}x - \frac{c}{m}v$$

undamped intuition

$$x(t) = A \cos(\omega t + \phi), \quad \omega = \sqrt{k/m}$$

The same structure appears in machines, suspensions, actuators, and vibrations.

Phase diagram



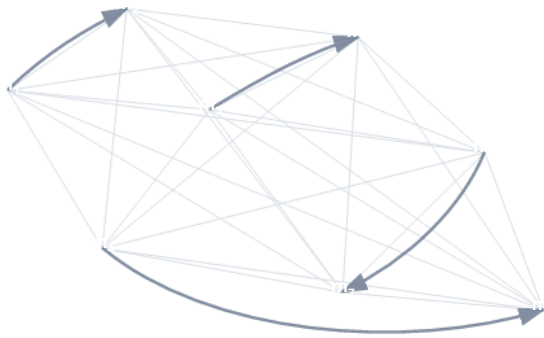
closed orbits: energy trades between x and v

State space makes the prediction visual: one point contains the system's future trajectory.

Simplicity = observable quantities + physical laws + initial conditions → predicted behavior

N-body dynamics: simple physical law + many interacting states -> complexity & emergent properties

Each mass pulls on every other mass: simple pairwise equations compound into high-dimensional “complex” behavior.



Every mass interacts with every other mass

N bodies → N positions + N velocities → a coupled trajectory in 6N-dimensional phase space

Many mass points, one repeated rule: sum the pairwise interactions.

Simple equations

position and velocity for each body

$$\dot{\mathbf{r}}_i = \mathbf{v}_i$$

pairwise Newtonian gravity

$$m_i \dot{\mathbf{v}}_i = \sum_{j \neq i} G m_i m_j \frac{\mathbf{r}_j - \mathbf{r}_i}{\|\mathbf{r}_j - \mathbf{r}_i\|^3}$$

state vector

$$\mathbf{y} = (\mathbf{r}_1, \dots, \mathbf{r}_N, \mathbf{v}_1, \dots, \mathbf{v}_N)$$

complexity grows fast

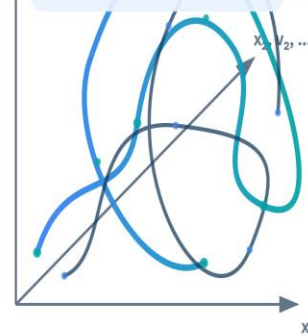
$$\dim(\text{phase space}) = 6N$$

Phase/state space

not one ellipse, but a high-dimensional landscape

The model is still compact

but the trajectory can become hard to visualize and sensitive



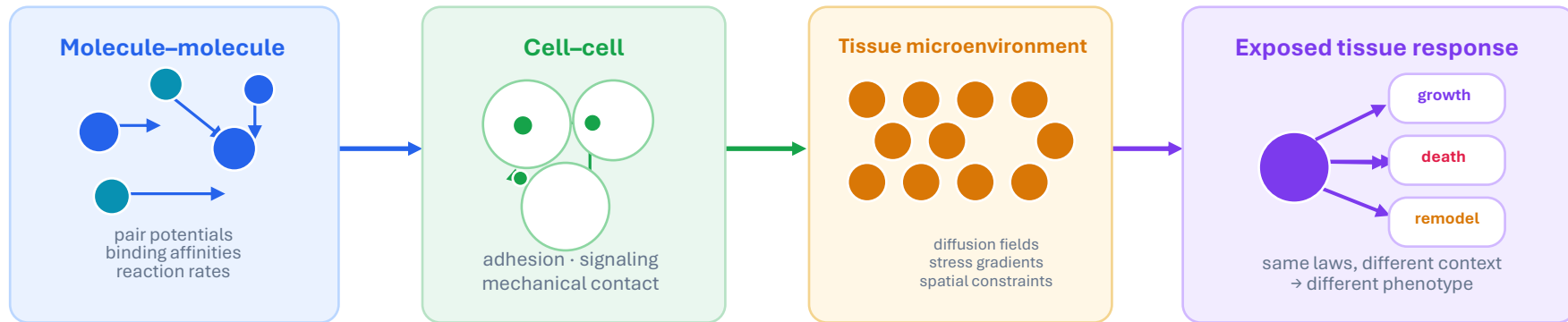
state = all positions + all velocities

The “point” in state space now lives in many dimensions; projections look tangled and sensitive.

Same philosophy: measure masses + initial states + apply one force law repeatedly → emergent, complex motion

Biological systems: simple micro-local rules, exponential system complexity

Known molecule–molecule interactions can be written compactly, but they propagate across scales into nonlinear, context-dependent tissue behavior.



Compact equations at each scale

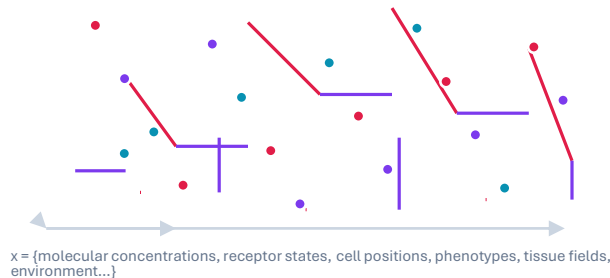
Molecules: ODEs to PDEs

Cells: ODEs to PDEs

Tissue: ODEs to PDEs

The notation stays concise — the coupled variables, feedback loops, and parameters multiply quickly.

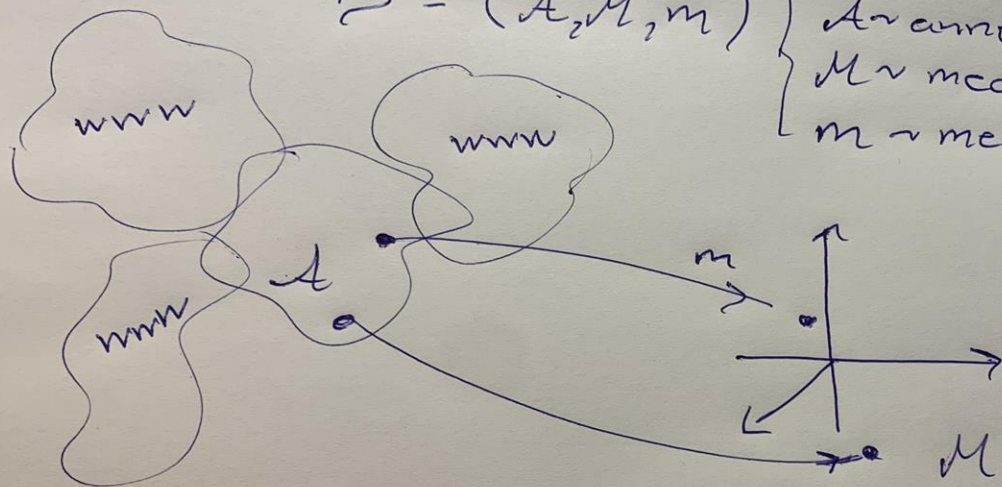
Phase / state space becomes huge and entangled



dimension $\gg 2$, nonlinear, multimodal, hierarchical, stochastic

A DATASET $\mathcal{D}$ IS A TRIPLE:

$$\mathcal{D} = (A, \mathcal{M}, m) \begin{cases} A \sim \text{annotation space} \\ \mathcal{M} \sim \text{measurement space} \\ m \sim \text{measurement map} \end{cases}$$

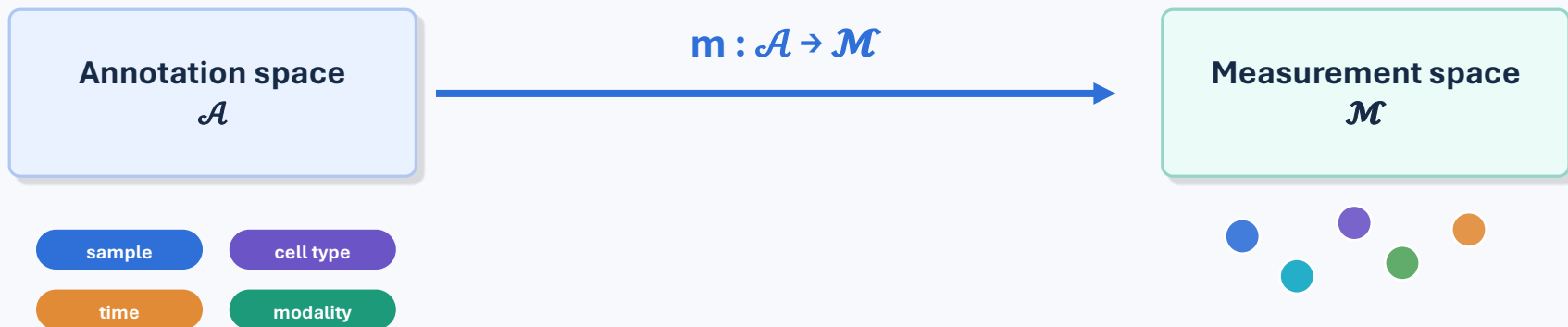


A is the realm of
AI agents

$\mathcal{M}$ is the realm
of more classical
math modeling
and statistical
learning

Annotated measurement maps

A mathematical language for data annotation and biomarker discovery



Data structures emerge when biological annotation is mapped into chosen geometries.

Mathematical modeling is performed on the triple $(A, M, m) \rightarrow$ new triples.

**For example; we create new annotations and we transform the measurement data in different ways;
the measurement map m is updated reflecting the modeling**

From annotation factors to measurement objects

Choosing which annotations remain external and which become internal changes the shape of the data.

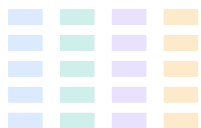
$$\mathcal{A} = S \times V$$

fixed sample s

→ vector $m_s \in \mathbb{R}^p$

all samples

→ matrix $X \in \mathbb{R}^{n \times p}$



$$\mathcal{A} = S \times T \times V$$

fixed sample s

→ trajectory $t \mapsto m(s, t) \in \mathbb{R}^p$

biomarkers may be
paths or temporal patterns



$$\mathcal{A} = S \times T \times V \times \text{Space}$$

fixed sample s and variable v

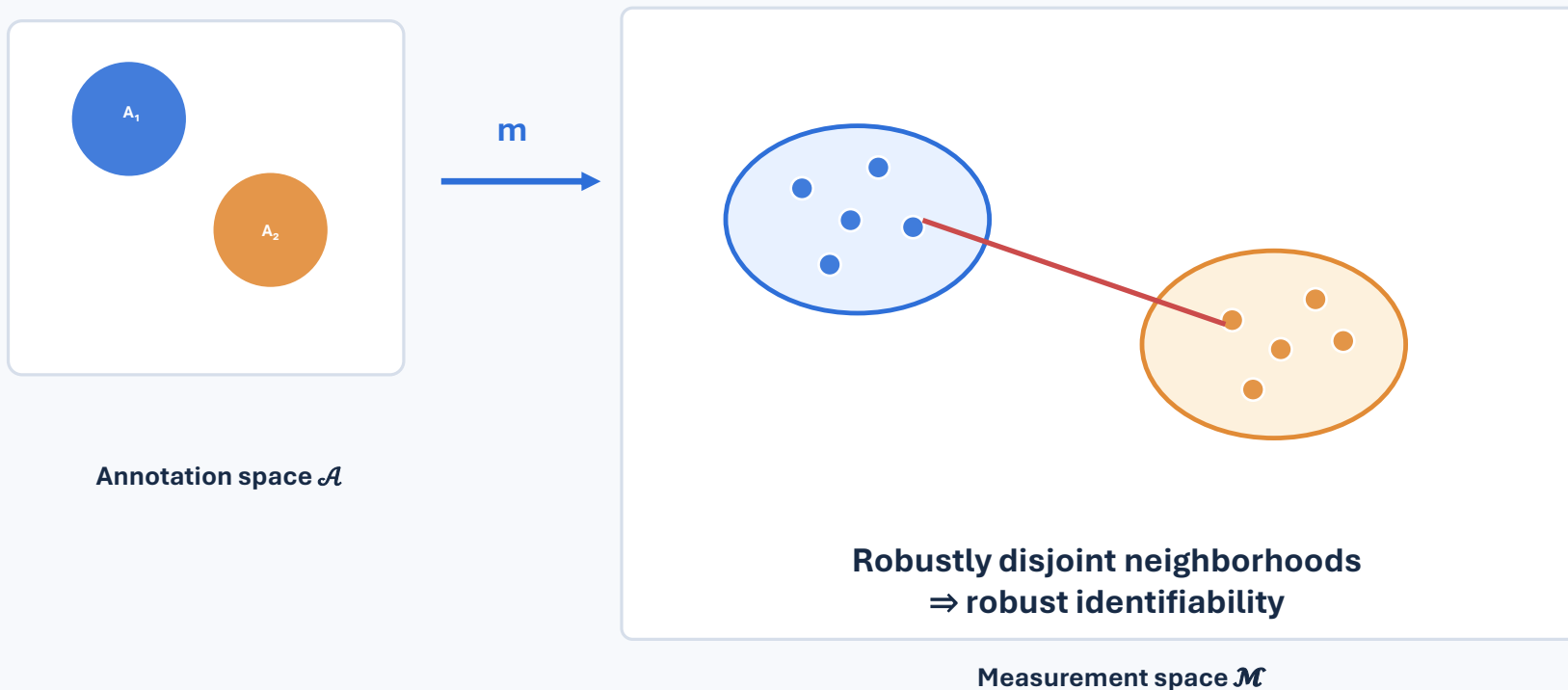
→ spatiotemporal field

biomarkers may be
shape, flow, or dynamics



Robust biomarkers as separation of the images of annotated domains

For disjoint domains A_1 and A_2 in annotation space, compare their images $X_1 = m(A_1)$, $X_2 = m(A_2)$ in measurement space.

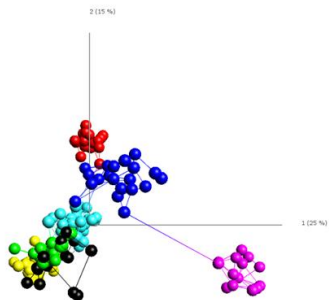


Geometry choices create different biomarker notions

The same annotation domains can become identifiable under e.g. graph distances, feature maps, or sparse coordinates.

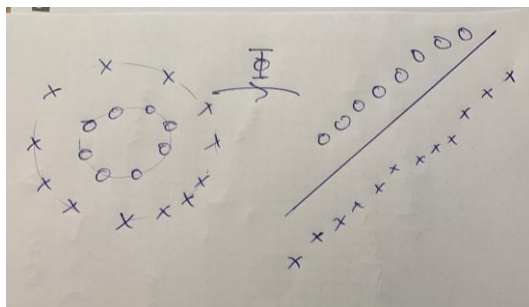
Graph geometry

Connect local neighbors; measure robust separation along paths.



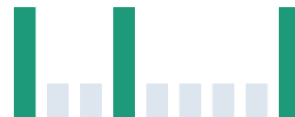
Kernel / feature space

A nonlinear map Φ can make classes linearly separable in $\mathcal{H}$.



Sparse signatures

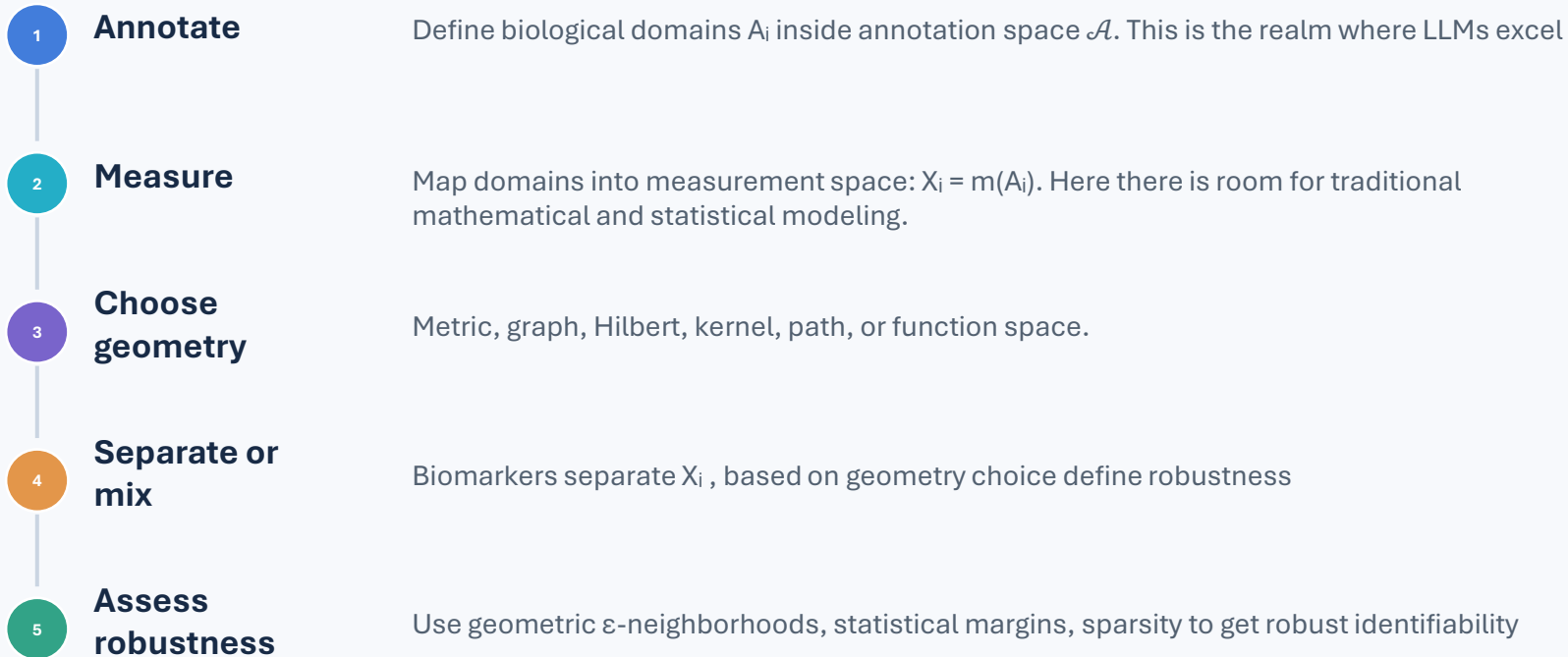
Weight vectors have small support; a few coordinates define the signature.



selected genes / features

Take-home synthesis

The same language unifies static biomarkers and dynamic biomarkers.



Biomarker discovery = robust, interpretable geometry of identifiable annotated measurement map domains.

Modeling and controlling immunodynamics

The Holy grail of modern biomedical research is to *predict and control* the human immune system

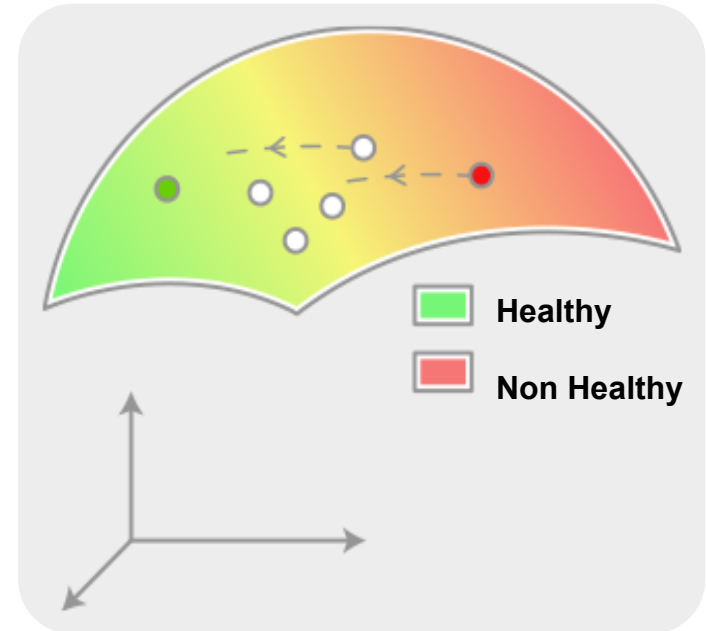


The immune system is central in

- Autoimmune disease
- Cancer
- Infectious disease
- Neurodegenerative disease
- And most other human disease conditions...

Homeostasis & Perturbations in the Immunity State Space

- *The Immunity State Space* is constructed from « deep/precision longitudinal and integrated measurements » on cells and other biomolecular actors
- Disease (autoimmune, neurodegenerative, cancer) are often connected with **low grade chronic inflammation**, i.e. a new **unfavorable stable attractor** in the state space
- The therapeutic goal is to, through interventions, push patients back to a **favorable (healthy) stable attractor**





***"If I have seen further, it is by standing on the shoulders of giants."* Isaac Newton 1675**

***"If I have seen further, it is by standing on the shoulders of giants."* Isaac Newton 1675**

Recent transformative developments in Biomedicine

Disruptive Discoveries & Inventions:

- Discovery of the DNA Double Helix structure (1953)
- Recombinant DNA and Protein Technologies (1970s onwards)
- Monoclonal Antibody Technology (1970s onwards)
- The Human Genome Project (completed 2003)
- CRISPR Gene Editing Technology (2012 onwards)
- Emergence of Gene and Cell Therapies

+ Proliferation of Advanced Measuring Technologies:

- New sequencing technologies (e.g., Next-Generation Sequencing)
- High-throughput measuring techniques (e.g., proteomics, metabolomics)

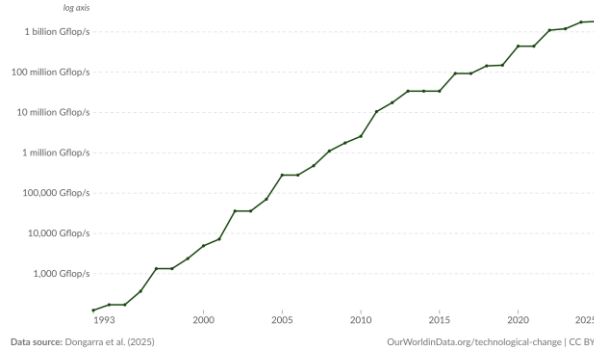
= **Result:** Generation of increasingly rich, detailed, and complex biomedical datasets, driving insights and opportunities, but with the potential for unprecedented acceleration...



Moore's law

Computational capacity of the fastest supercomputers

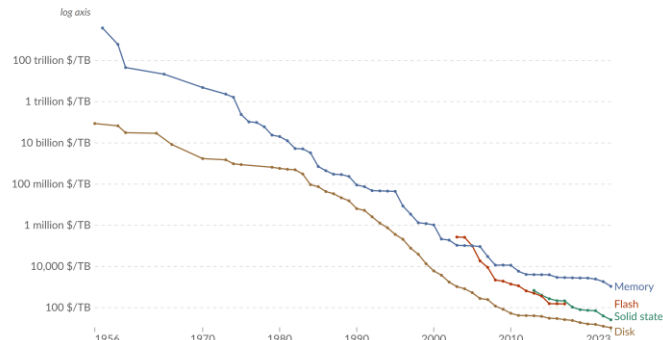
The number of floating-point operations¹ carried out per second by the fastest supercomputer in any given year. This is expressed in gigaflops, equivalent to 10⁹ floating-point operations per second.



1. Floating-point operation A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

Historical price of computer memory and storage

This data is expressed in US dollars per terabyte (TB), adjusted for inflation. "Memory" refers to random access memory (RAM), "disk" to magnetic storage, "flash" to special memory used for rapid data access and rewriting, and "solid state" to solid-state drives (SSDs).



Data source: John C. McCallum (2023); U.S. Bureau of Labor Statistics (2026)

OurWorldInData.org/technological-change | CC BY

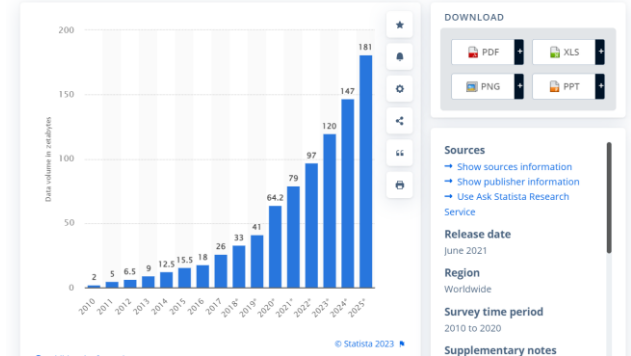
Note: For each year, the time series shows the cheapest historical price recorded until that year. This data is expressed in constant 2020 US\$.

Technology & Telecommunications > Telecommunications

PREMIUM

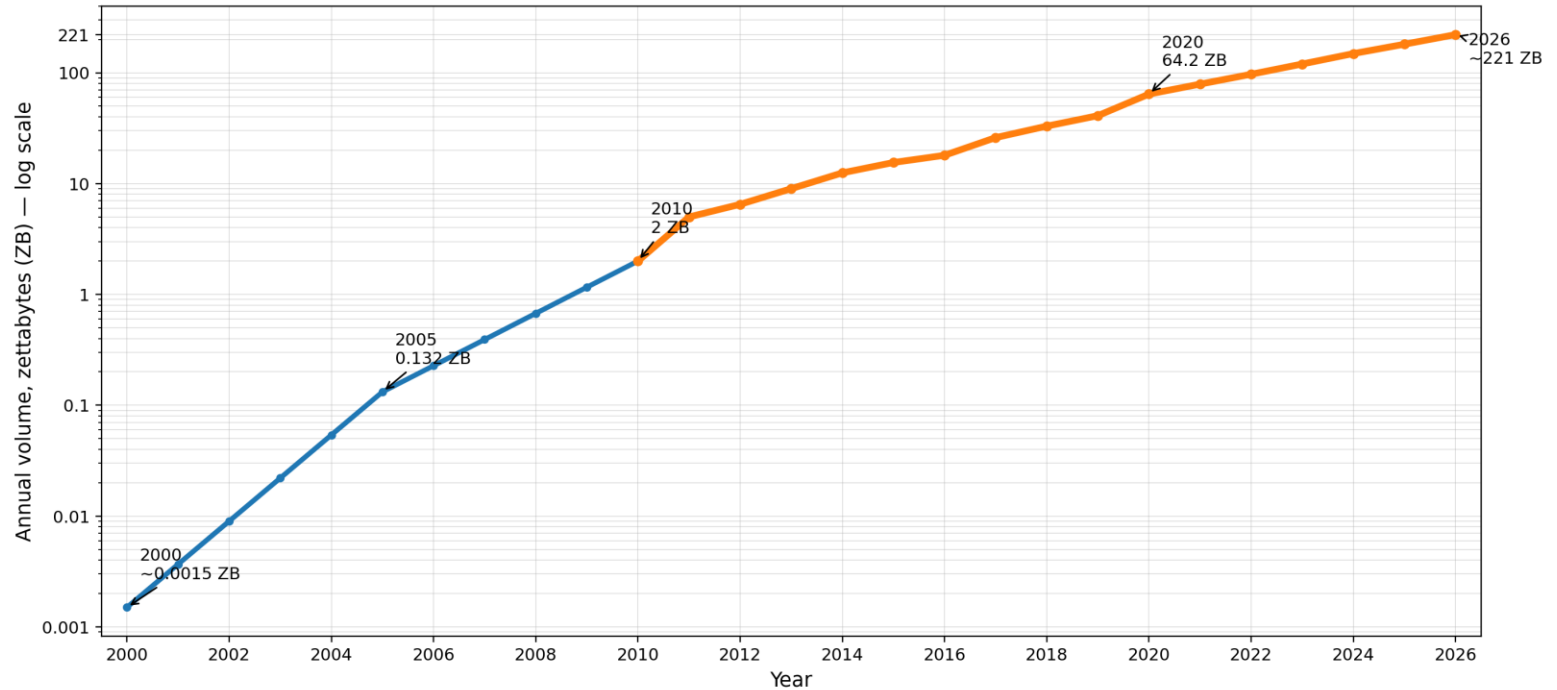
Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

(in zettabytes)



Exp growth of Data volumes & Computational capacity & Data storage per dollar

Worldwide Data Created, Captured, Copied & Consumed (2000–2026)



Notes: 2000–2009 are anchor-based estimates/interpolations from earlier 'digital universe' measurements.
2010 onward follows the modern Statista/IDC-style Global DataSphere definition; 2021–2026 are forecasts/projections.

Category

All global data

Healthcare / biomedical proxy at 30%

Healthcare / biomedical proxy per year

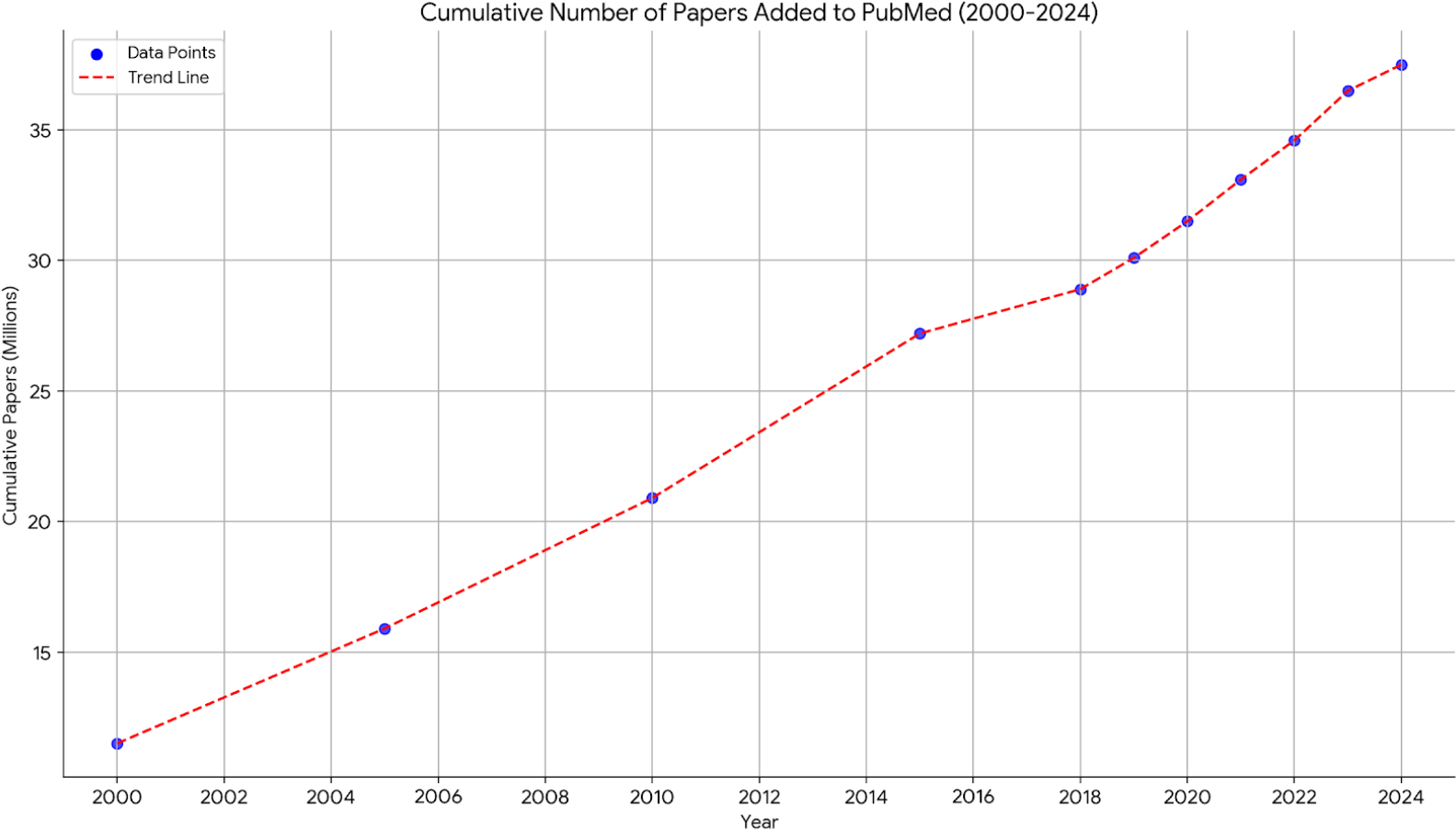
Approx. rate now

~7.0 PB/second

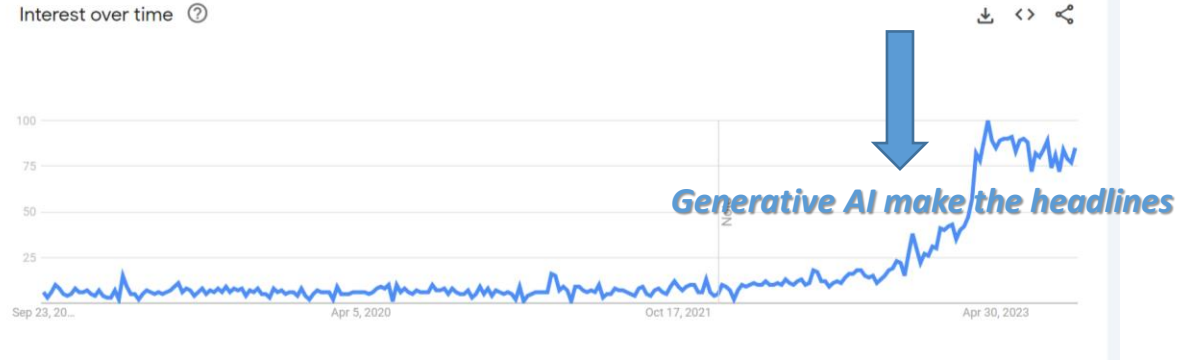
~2.1 PB/second

~66 ZB/year

PubMed database currently grows at around 3 articles per minute



And AI... A Google Trends search on *Artificial Intelligence*



Please comment on the Human Machine partnership



Thought for a few seconds >

A strong **Human–Machine partnership** is not about replacing people with technology; it is about designing work so that humans and machines each do what they are best suited for.

Machines are excellent at scale, speed, consistency, pattern detection, simulation, monitoring, and handling repetitive tasks. Humans remain essential for judgment, context, ethics, empathy, creativity, accountability, and deciding what outcomes actually matter.



Top Takeaways

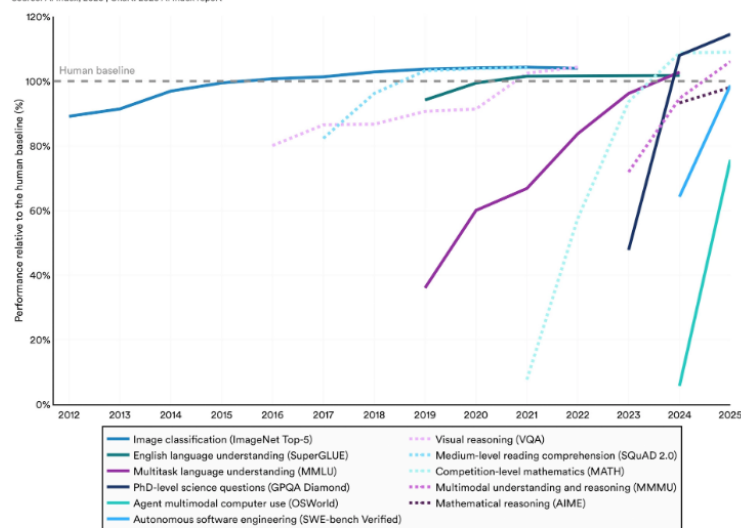
1. AI capability is not plateauing. It is accelerating and reaching more people than ever.

Industry produced over 90% of notable frontier models in 2025, and several of those models now meet or exceed human baselines on PhD-level science questions, multimodal reasoning, and competition mathematics. On a key coding benchmark—SWE-bench Verified—performance rose from 60% to near 100% in a single year. Organizational adoption reached 88%, and 4 in 5 university students now use generative AI.

See: [Chapter 2: Technical Performance](#)

Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2026 | Chart: 2026 AI Index report



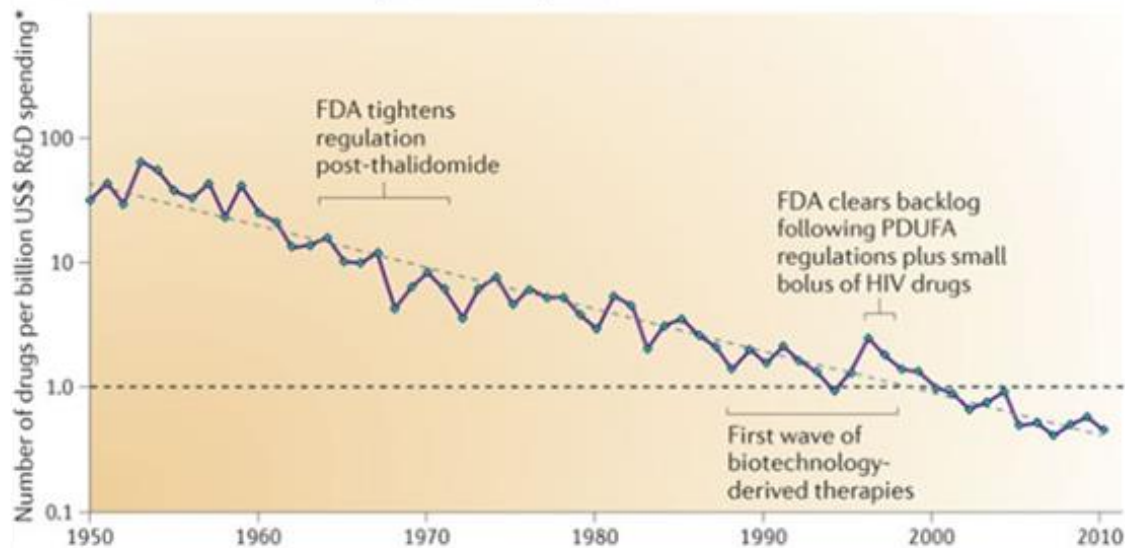
Infrastructure, data, compute and AI is growing exponentially BUT

Eroom's law for cost in drug development

Scannell, J., Blanckley, A., Boldon, H. et al. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat Rev Drug Discov* 11, 191–200 (2012).

<https://doi.org/10.1038/nrd3681>

a Overall trend in R&D efficiency (inflation-adjusted)





Feature

The pharmaceutical productivity gap – Incremental decline in R&D efficiency despite transient improvements

Kenneth D.S. Fernald  , Philipp C. Förster, Eric Claassen, Linda H.M. van de Burgwal

[Show more](#) 

 Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.drudis.2024.104160> 

[Get rights and content](#) 

Under a [Creative Commons license](#) 

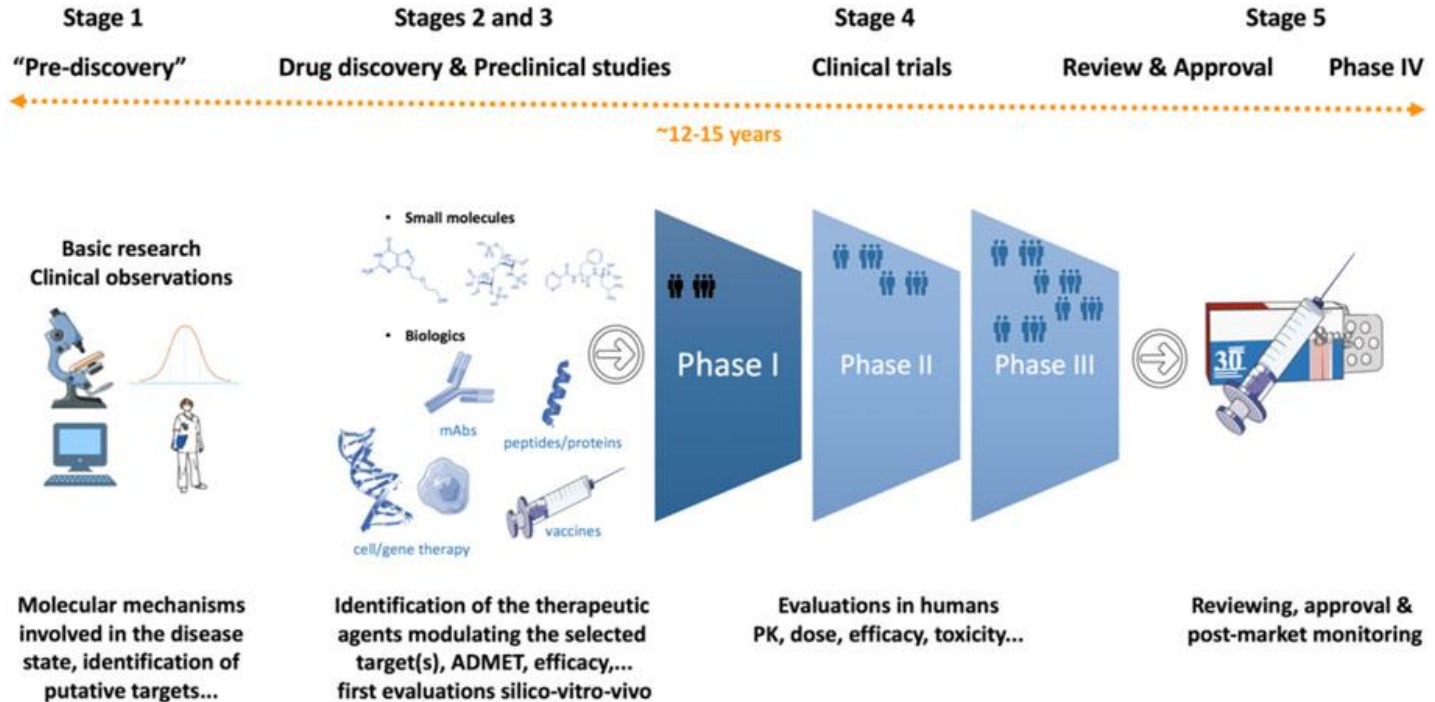
 [Open access](#)

Highlights

- Rising R&D expenditures reflect ongoing incremental efficiency decline.
- Recent stabilization in R&D efficiency may not indicate structural reversal.
- Sustained turnaround in R&D efficiency remains uncertain.
- Rising late-stage attrition signals lower future R&D productivity.
- Structural changes are essential to address pharmaceutical innovation challenges.

The approximate cost to put a molecule on the market is around US \$3.5 billion US dollars and the time needed to complete the R&D process is around 12–15 years.

Drug discovery and development

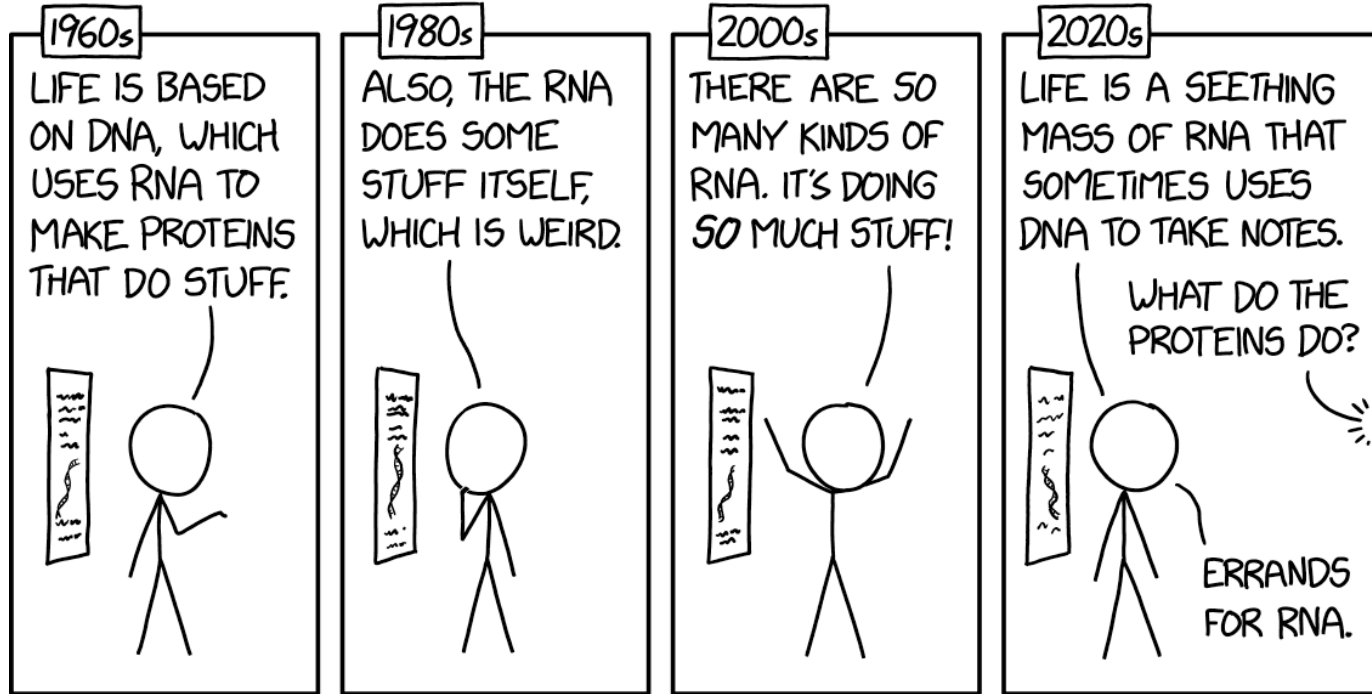


Multiple factors might explain this lack of direct positive correlation between the underlying capabilities and resources in biomedicine and biotechnology on the one hand, and the capacity to deliver tangible results for patients on the other hand.

- **Better than the Beatles' problem:** An ever-growing catalogue of existing approved drugs continuously raises the bar for the development of new drugs
- **The “cautious regulator” problem:** An ever-lowering of the risk tolerance of drug regulatory agencies raises the bar for the approval of new drugs
- **The “throw money at it” tendency:** An ever-increasing tendency to, without introducing fundamental structural changes in the R&D processes, increase resources to R&D, which not necessarily correlates linearly with output
- **The 'basic research–brute force' bias:** The 'basic research–brute force' bias is the tendency to overestimate the ability of advances in basic research, and brute force screening methods to increase the probability that a molecule will be safe and effective

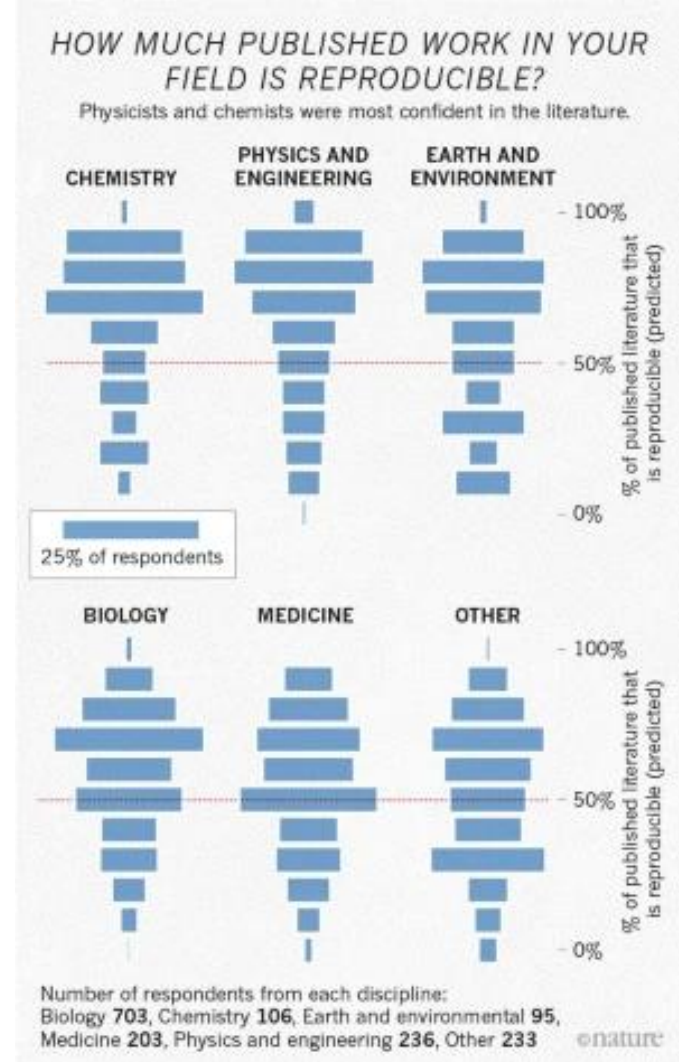
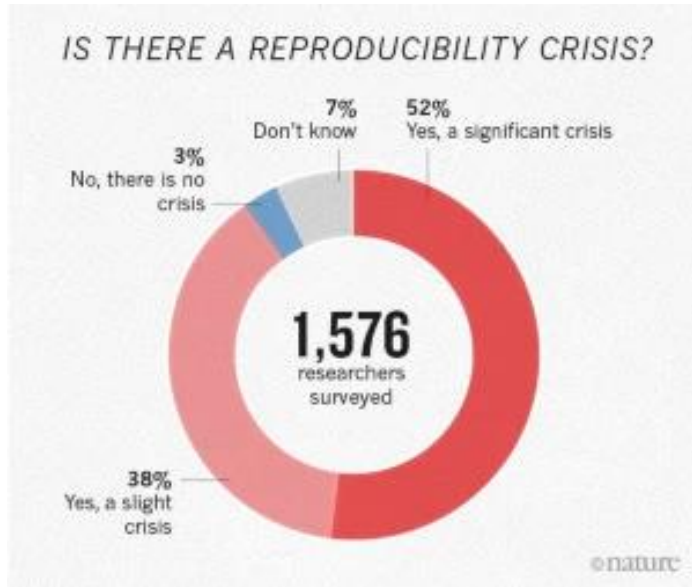
Scannell, J., Blanckley, A., Boldon, H. et al. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov 11, 191–200 (2012). <https://doi.org/10.1038/nrd3681>

BIOLOGY is complex and understanding takes time



Biological models are only partly verifiable

Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
<https://doi.org/10.1038/533452a>



Biomedical researchers' perspectives on the reproducibility of research

Kelly D. Cobey , Sanam Ebrahimzadeh, Matthew J. Page, Robert T. Thibault, Phi-Yen Nguyen, Farah Abu-Dalfa, David Moher

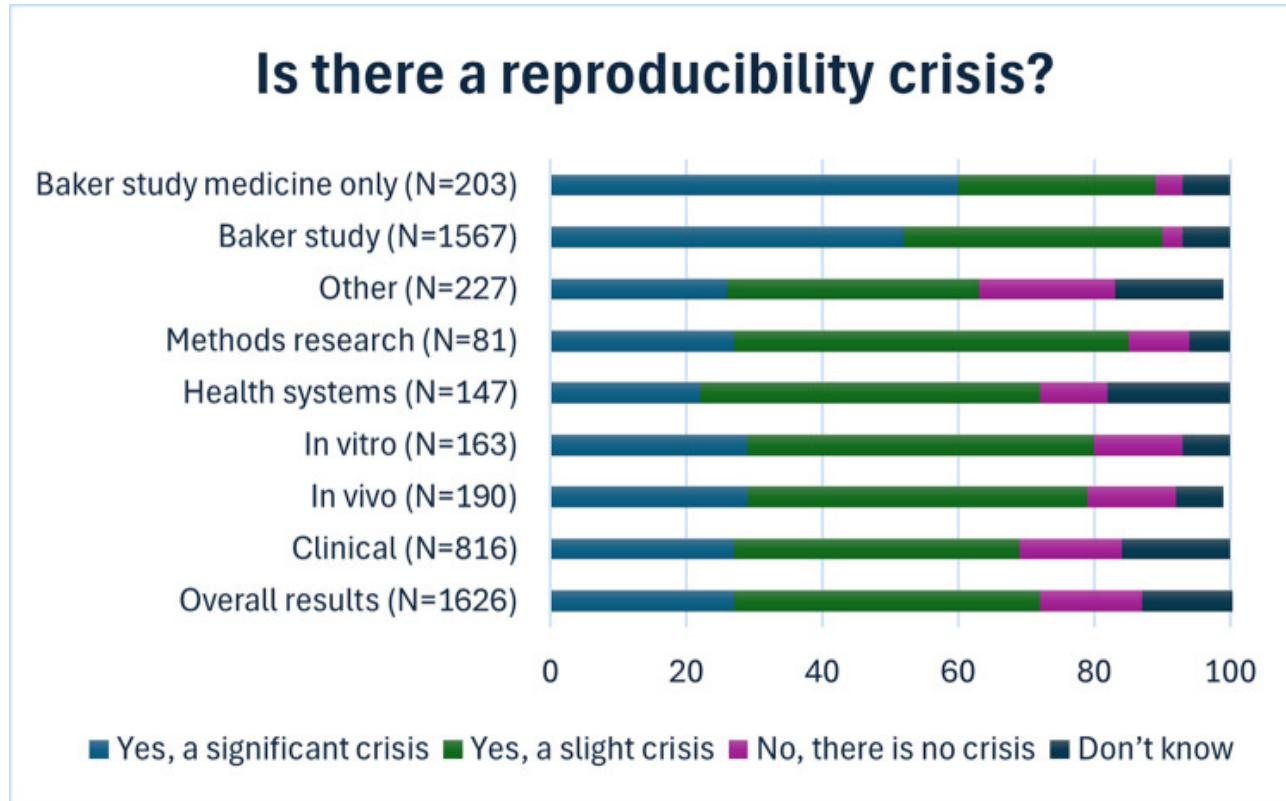
Published: November 5, 2024 • <https://doi.org/10.1371/journal.pbio.3002870>

Article	Authors	Metrics	Comments	Media Coverage	Peer Review
					

“We report the results of an international survey examining perceptions of reproducibility. Almost 3 quarters of participants reported that they felt there was a reproducibility crisis in biomedicine. The concern appears to apply to biomedicine overall, but also specifically to clinical research, in vivo research, and in vitro research (11% or fewer participants indicated that they think more than 80% of papers in each category were reproducible).”

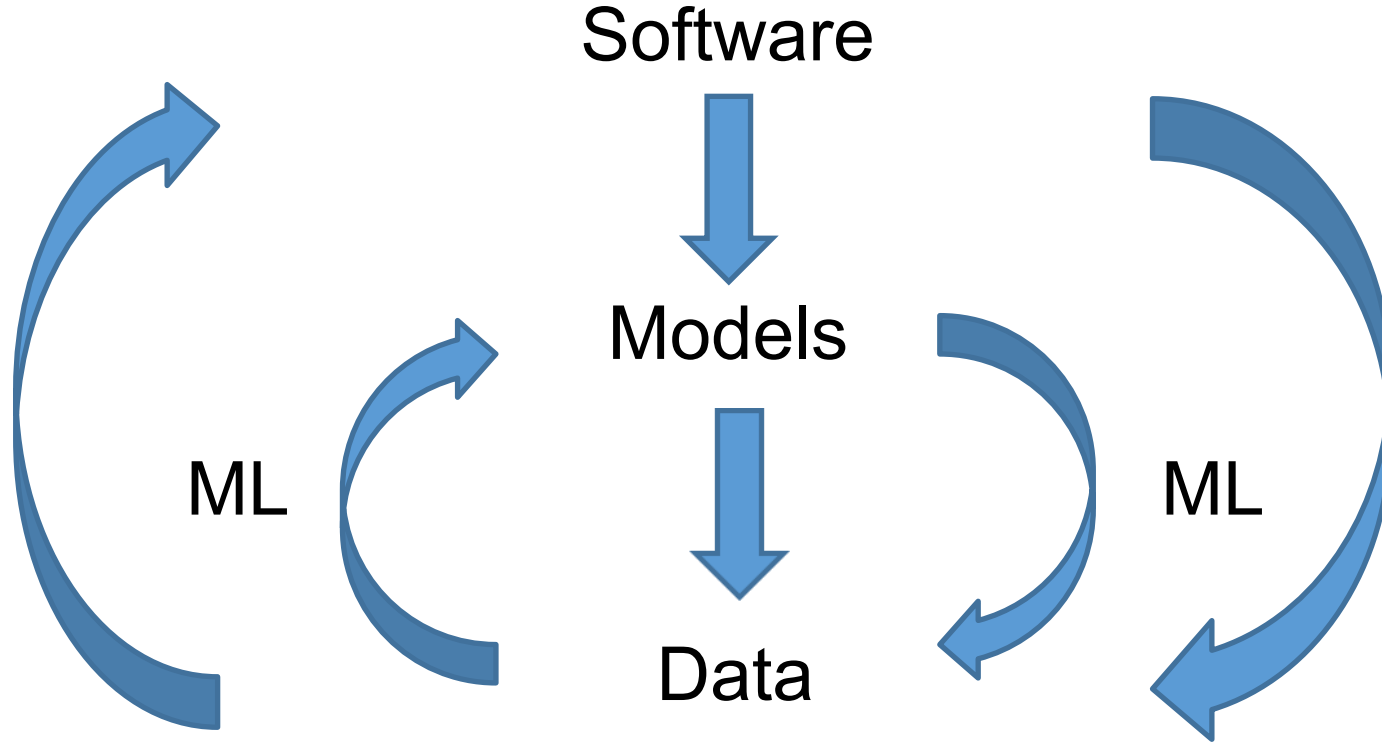
Biomedical researchers' perspectives on the reproducibility of research

PLoS Biol. 2024 Nov 5;22(11):e3002870. doi: 10.1371/journal.pbio.3002870



What is AI ?

Traditional vs AI- Machine Learning Computational Analysis

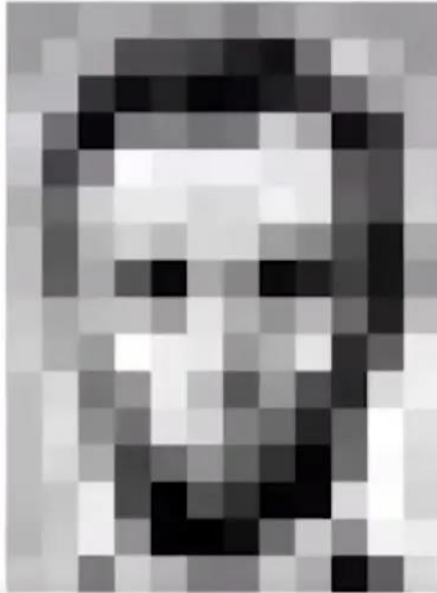


Deep Learning (Convolutional Neural Networks)) was first successful in Image Analysis

Transformer : Translating light intensities in pixels to numbers

DIGITAL
IMAGES

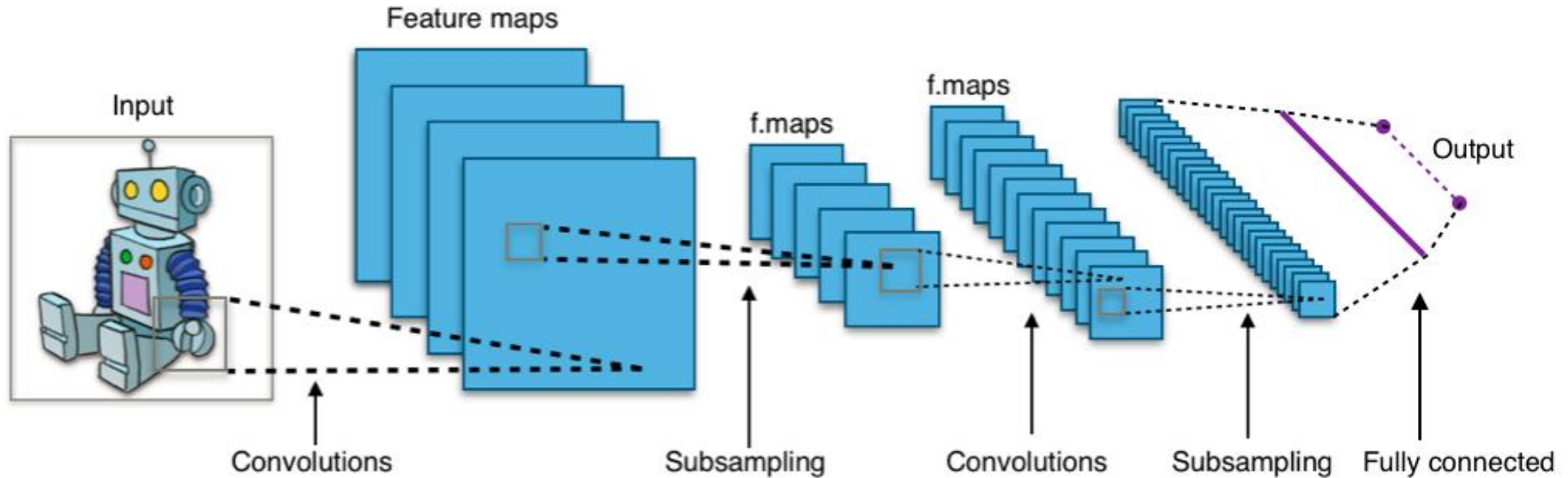
BLACK: 0 - - - - - WHITE: 255



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	106	207	233	233	214	230	239	228	98	74	206
188	88	179	209	180	215	211	168	138	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	188	227	178	143	182	106	96	190
205	174	155	252	236	231	149	179	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

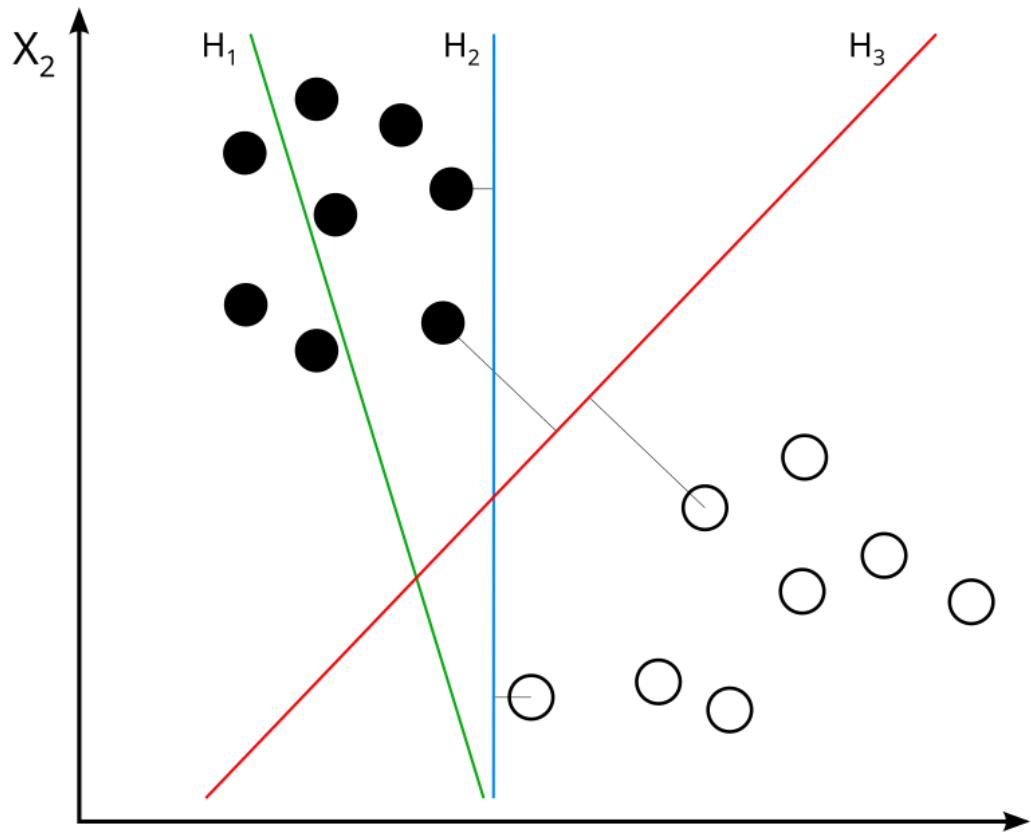
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	106	207	233	233	214	230	239	228	98	74	206
188	88	179	209	180	215	211	168	138	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	188	227	178	143	182	106	96	190
205	174	155	252	236	231	149	179	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Expanding the variable (feature) space

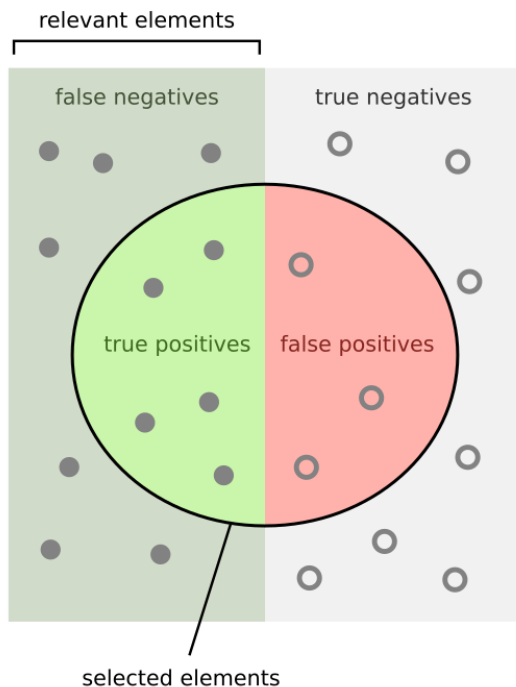


And apply traditional Machine Learning like Support Vector Machines (SVMs)

An example of how to design a Predictor



Basic characteristics for Predictors



How many relevant items are selected?
e.g. How many sick people are correctly identified as having the condition.

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

How many negative selected elements are truly negative?
e.g. How many healthy people are identified as not having the condition.

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

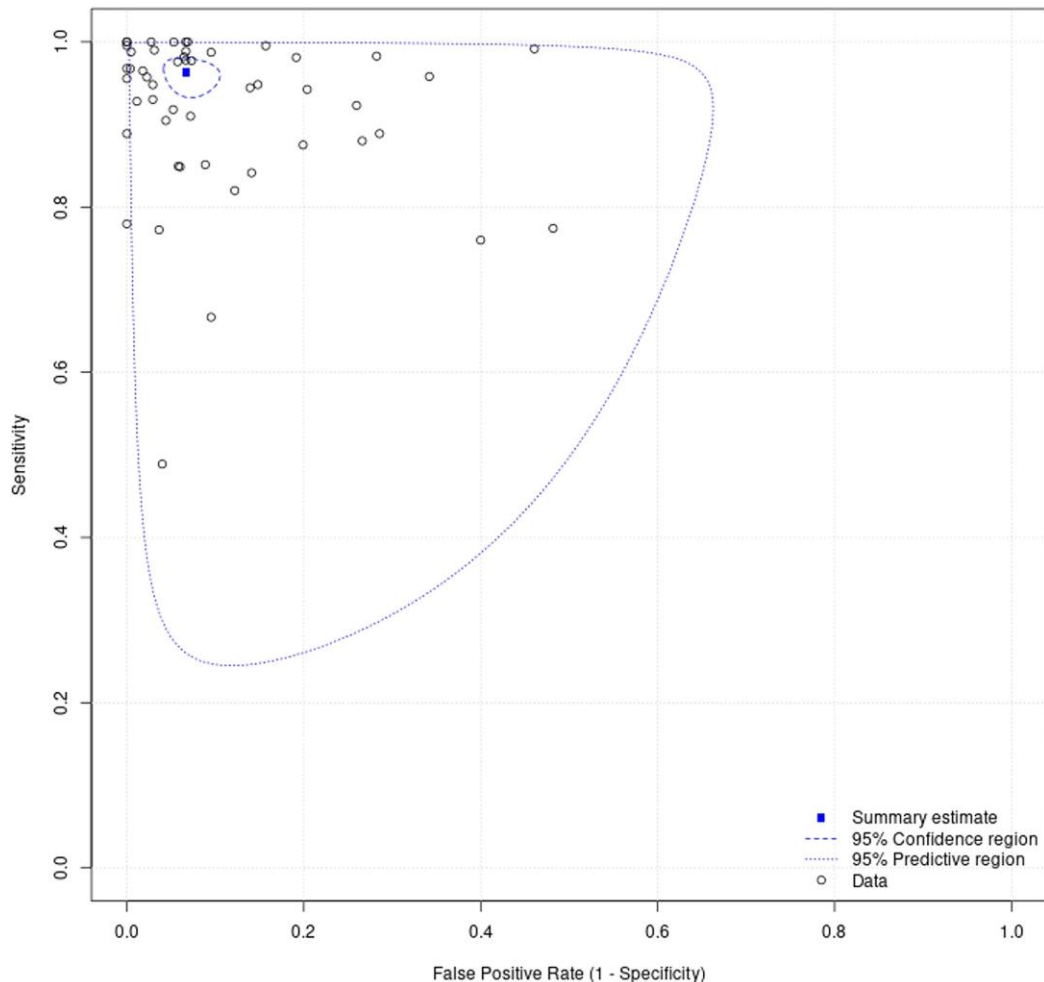
ML image analysis is already quite mature

Summary receiver operating characteristic plot of AI applied to whole slide images for all disease types generated from MetaDTA: diagnostic test accuracy meta-analysis v2.01 Shiny App
https://crsu.shinyapps.io/dta_ma/

Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy

Clare McGenity, Emily L. Clarke, Charlotte Jennings, Gillian Matthews, Caroline Cartledge, Henschel Freduah-Agyemang, Deborah D. Stocken & Darren Treanor

npj Digital Medicine volume 7, Article number: 114 (2024) Cite this article

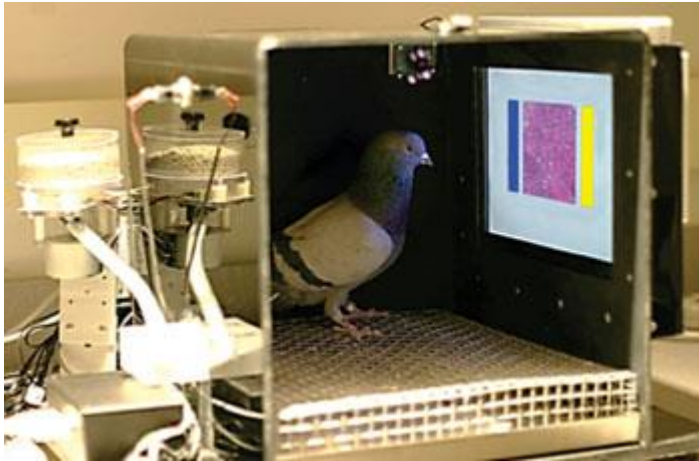


The Human Baseline

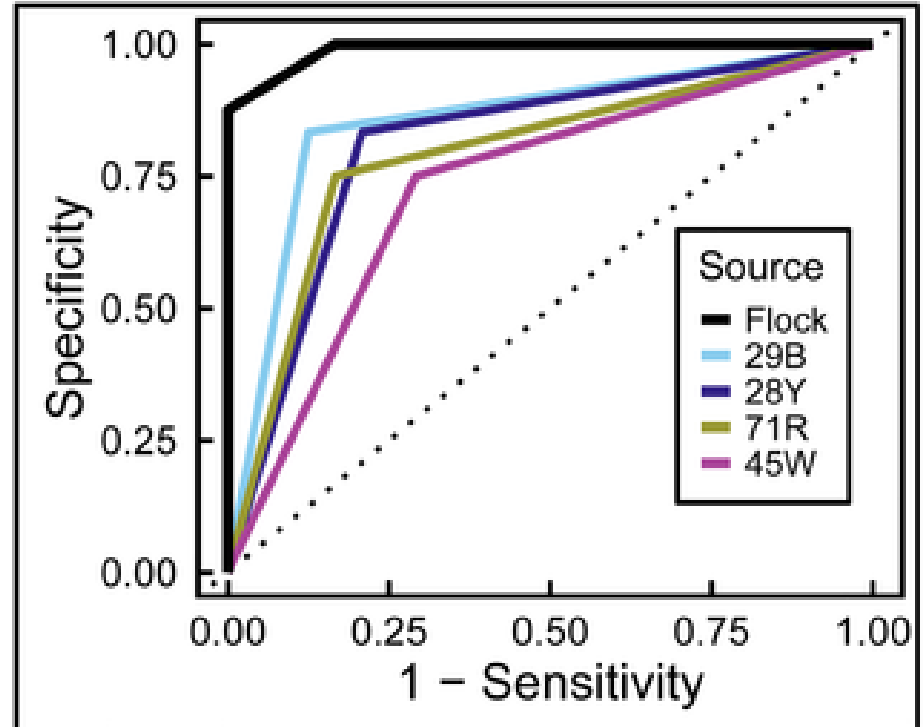
Overall accuracy, concordance rates (above- vs below-consensus), mean under versus over-calling rates, and review time (in seconds) as a function of consensus diagnosis. Standard deviations included in parentheses.

Consensus Diagnosis	Mean (95%CI) Concordance Rate	Above Consensus Diagnosis (95% CI)	Below Consensus Diagnosis (95% CI)	Mean (95%CI) Under/Over-calling Rate	Mean (95%CI) Review Time (sec)
Benign	71% (61%, 82%)	29% (20%, 40%)	-	.29 (.18, .40)	120.6 (101.6, 139.5)
Atypia	37% (29%, 45%)	21% (15%, 28%)	43% (35%, 50%)	-.22 (-.35, -.09)	193.9 (140.4, 247.5)
DCIS	52% (43%, 61%)	17% (12%, 23%)	31% (25%, 39%)	-.14 (-.25, -.04)	250.4 (164.8, 336.1)
Invasive	94% (88%, 99%)	-	6% (2%, 14%)	-.06 (-.12, -.01)	205.6 (147.9, 263.3)

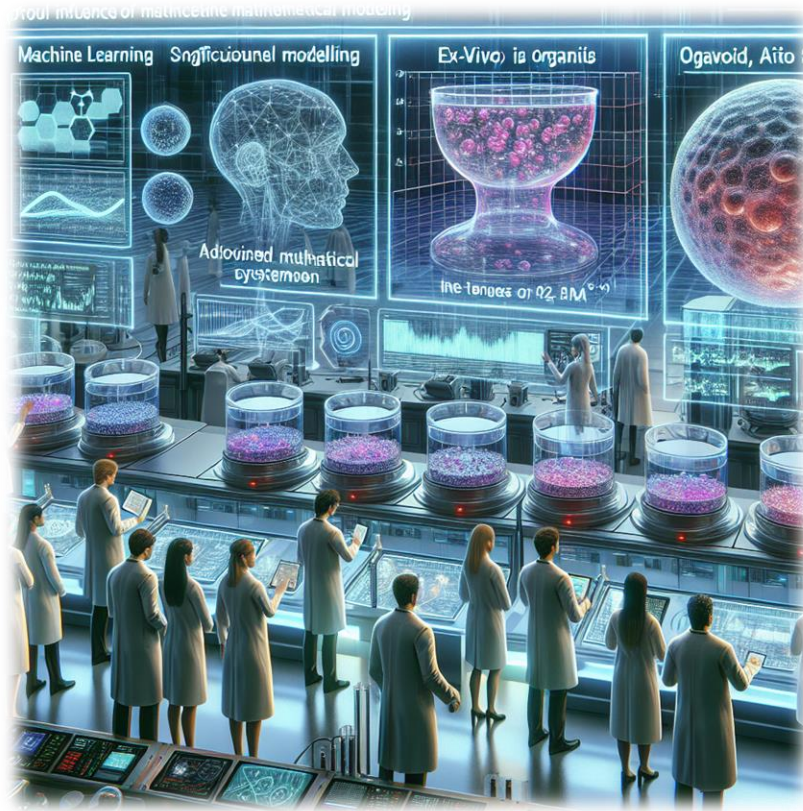
And how good are Pigeons?



Pigeons (*Columba livia*) as Trainable Observers of
Pathology and Radiology Breast Cancer Images
Richard M. Levenson ,Elizabeth A. Krupinski,Victor M.
Navarro,Edward A. Wasserman
Published: November 18, 2015
<https://doi.org/10.1371/journal.pone.0141357>



It is hard to predict, especially the future



***“A chess problem is genuine mathematics, but it is in some way 'trivial' mathematics”.
From A Mathematician's Apology, a 1940 essay by the British mathematician G. H. Hardy***

Chess as a toy-model for thinking about the future of Science

In Chess, Human vs Machine: In 1997 it was GAME OVER



*Each of Deep Blue's chips consisted of 1.5 million transistors
.....Together, the chips could churn through 200 million chess positions
per second. **That brute-force power, combined with clever game-
evaluation functions, gave Deep Blue decisive moves that Kasparov
called “uncomputerlike.”***

<https://spectrum.ieee.org/chip-hall-of-fame-ibm-deep-blue-2-chess-chip>



Kasparov contemplating the Human-Machine Partnership

Kasparov's law:

"A weak human + machine + better process was superior to a strong computer alone and ... superior to a strong human + machine + inferior process."

"Machines have calculations. We have understanding. Machines have instructions. We have purpose. Machines have objectivity. We have passion....

....There is one thing that only a human can do. That's to dream. So let us dream big."

<https://www.kasparov.com/deep-thinking-ai/>

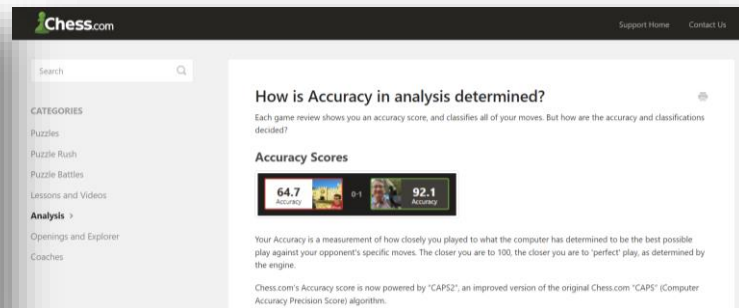
Take away: Focus on processes + interfaces...

Today, *a computer-like move*, means highest praise in chess...



**Hikaru Nakamura + Magnus Carlsen lose
100% of their chess games vs Stockfish**

<https://www.youtube.com/watch?v=OT-1sG-sZLU>

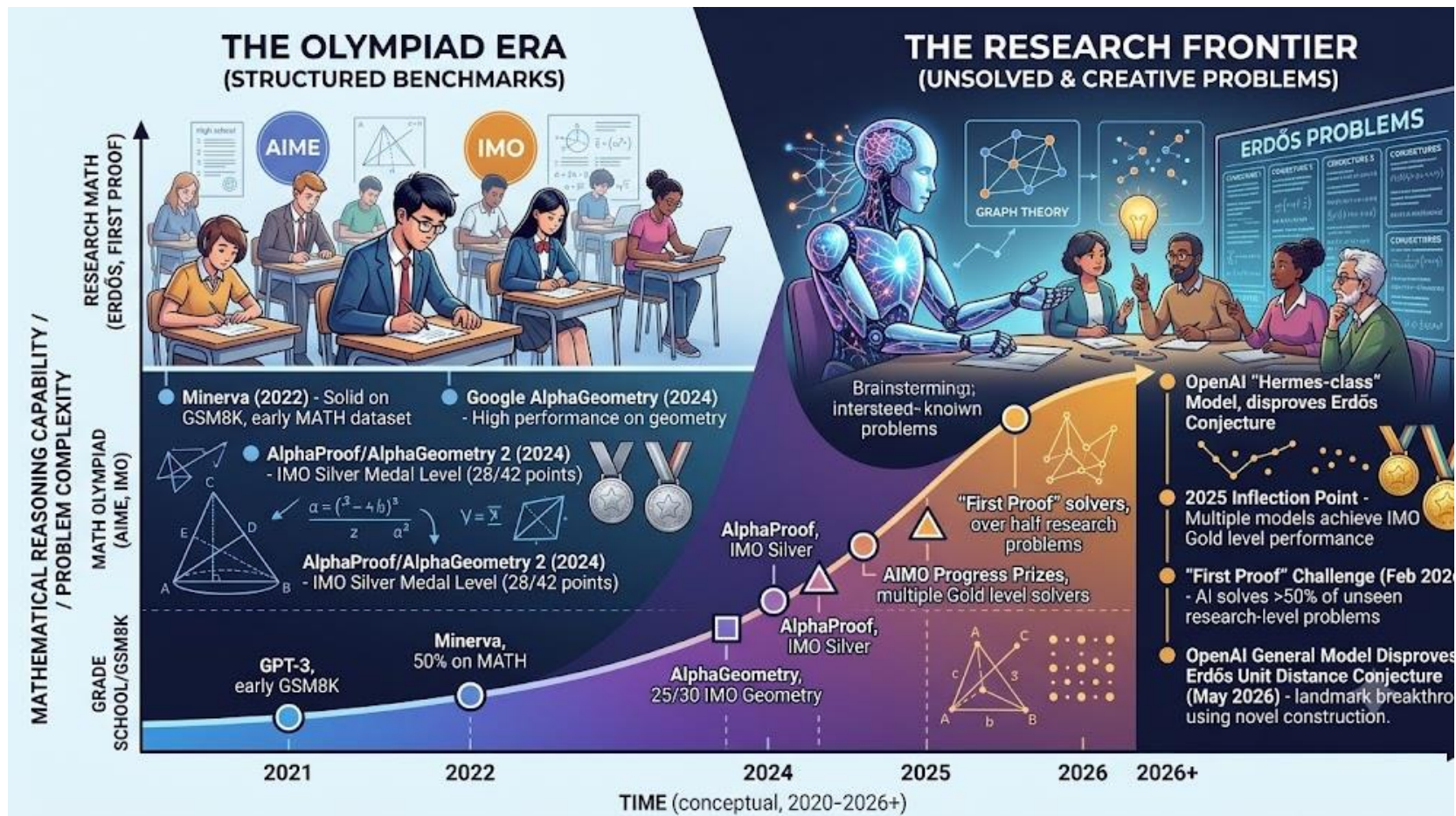


- ***No human chess player today stands a chance against a reasonable chess engine***
- ***Accuracy in chess is estimated based on best play as assessed by machines***
- ***Training and analysis in chess are heavily machine dependent***

Rating Stockfish on a human scale (e.g. FIDE Elo) has become an almost impossible task, as the difference in strength between it and humans is now so large that this difference can hardly be measured.

<https://official-stockfish.github.io/docs/stockfish-wiki>

Machines doing Mathematics. Some hallucinations from Gemini:



Where did Gemini find the data for the "First Proof" challenge?

● SECOND BATCH ANNOUNCEMENT

[This document \(PDF\)](#) describes our plan for a second batch of problems, which will be created, tested, and graded from March to June 2026. This batch will be designed as a formal **benchmark**. It will also include a separate round of informal **community experimentation**, in which a set of problems is made available to the interested public, with solutions provided after a few days, followed by an open discussion.

○ TIMELINE

Problem selection

March — May 2026

Mathematicians across fields submit unpublished problems with proofs of at most 8 pages. All submissions undergo a first round of refereeing, and 10 problems are selected for the benchmark plus a separate set for community experimentation.

Benchmark testing

Late May — Early June 2026

The editorial board tests AI systems via API. Each system is an open source harness which calls publicly available models, and gets one shot per question with no additional interaction.

Benchmark grading

June 2026

Human mathematicians referee AI solutions blind, rating each as essentially flawless, publishable with minor revisions, requiring major revisions, or rejected. Results — including referee reports and editorial reasoning — will be published online.

● RESULTS

To be announced on June 10, 2026.

ArXivMath: Evaluating LLMs on Mathematical Research Problems From Recent ArXiv Papers

February 04, 2026 | By [Jasper Dekoninck](#), [Tim Gehringer](#), [Martin Vechev](#) | [Discuss on X](#)



This work was partially supported by an evaluation grant from Google.



MathArena

🎉 **New (June 4):** We added [Step 3.7 Flash](#) to the leaderboard!

🎉 **New (June 2):** We added [Claude Opus 4.8](#) to the leaderboard!

[ArXivLean](#)[BrokenArxiv](#)[ArXivMath](#)[Visual Math](#)[Final-Answer Comps](#)[Proof-Based Comps](#)[Project Euler](#)

Overall

12/2025

01/2026

02/2026

03/2026

04/2026



Detailed View

Rank	Model Name	Provider	Accuracy (± 95% CI)	Cost	Output Tokens	Input Tokens	Average Retries	Average Time	
1	GPT-5.5 (xhigh) 🚩	OpenAI	67.07% ± 10.17%	\$0.62		20665	936	0.00	6m 15s
2	Gemini 3.1 Pro Preview	Google	62.20% ± 10.50%	\$0.32		26392	171	0.00	3m 37s
3	Claude-Opus-4.8 (max) 🚩	Anthropic	60.98% ± 8.62%	\$3.45	138012	292	0.00	0m 21s	
4	Claude-Opus-4.7 (xhigh) 🚩	Anthropic	58.54% ± 10.66%	\$0.87		34898	294	0.00	0m 0s
5	DeepSeek-v4-Pro (Max) 🚩	DeepSeek	52.44% ± 10.81%	\$0.33	96046	236	0.20	50m 56s	
6	Gemini 3.5 Flash 🚩	Google	51.22% ± 8.83%	\$0.22		24812	169	0.00	1m 44s
7	DeepSeek-v4-Flash (Max) 🚩	DeepSeek	48.78% ± 10.82%	\$0.037	133813	245	0.06	27m 6s	
8	Step 3.7 Flash 🚩	StepFun	37.40% ± 8.55%	\$0.042	141372	198	0.00	20m 32s	
9	GLM 5.1 🚩	Z.ai	36.59% ± 10.43%	\$0.15		48201	181	0.01	17m 41s
10	Step 3.5 Flash	StepFun	34.15% ± 10.26%	\$0.030	99434	227	0.00	12m 3s	
N/A	Qwen3.5-2B	Qwen	N/A	N/A		51175	226	0.00	127m 8s

May 20, 2026 [Research](#) [Milestone](#)

An OpenAI model has disproved a central conjecture in discrete geometry

[Read the proof](#)[Read the companion remarks](#)

REMARKS ON THE DISPROOF OF THE UNIT DISTANCE CONJECTURE

NOGA ALON, THOMAS F. BLOOM, W. T. GOWERS, DANIEL LITT, WILL SAWIN, ARUL SHANKAR,
JACOB TSIMERMAN, VICTOR WANG, AND MELANIE MATCHETT WOOD

ABSTRACT. We present a short, digested, human-verified version of the recent OpenAI-generated counterexample to the Erdős unit distance conjecture, and a sequence of reflections on it. The argument relies crucially on ideas that may, at least in retrospect, be attributed to Ellenberg-Venkatesh, Golod-Shafarevich, and Hajir-Maire-Ramakrishna.

CONTENTS		5. W T Gowers	9	10. Victor Wang	16
		6. Daniel Litt	12	11. Melanie Matchett Wood	17
1. Introduction	1	7. Will Sawin	13	References	17
2. Proof and further overview	3	8. Arul Shankar	15		
3. Noga Alon	7	9. Jacob Tsimerman	16		
4. Thomas Bloom	7				

1. INTRODUCTION

The following result is due to an internal model at OpenAI. The proof we give in these remarks is a human-digested, somewhat simplified, and somewhat generalized version of the AI proof.¹

¹Throughout this document, phrases such as “AI proof” or “GPT proof” all refer to the same file, which was first mathematically generated in one shot by an internal model at OpenAI, and then expositively refined through human interactions with Codex.

Leiden Declaration on Artificial Intelligence and Mathematics

2 June 2026 · [DOI: 10.5281/zenodo.20302944](https://doi.org/10.5281/zenodo.20302944)

Preamble

Technological developments have repeatedly transformed the practice of mathematics. Recent artificial intelligence technologies, including symbolic and neural methods for the generation and formalization of mathematics, may already have initiated a significant chapter in this long history. Among researchers, artificial intelligence has produced a wide range of reactions: enthusiasm for its potential to yield new discoveries; intimidation by the pace of developments; indifference to these rapid changes; and concern for the implications, both for mathematics and in wider society.

Mathematicians have a choice about whether and how to adopt artificial intelligence in the conduct of their research. They also have a responsibility to ensure the continued flourishing of the discipline. This Declaration calls upon mathematicians to exercise this responsibility, and provides recommendations for individuals, institutions, government, and industry.

Mathematics as the language of Human Thought

Philosophy [i.e. natural philosophy] is written in this grand book—I mean the Universe—which stands continually open to our gaze, but it cannot be understood unless one first learns to comprehend the language and interpret the characters in which it is written. It is written in the language of mathematics, and its characters are triangles, circles, and other geometrical figures, without which it is humanly impossible to understand a single word of it; without these, one is wandering around in a dark labyrinth.



Translated from Il saggiatore by Galileo Galilei, published in Rome in October 1623.
Presenting the Scientific Method

The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure, even though perhaps also to our bafflement, to wide branches of learning.

From THE UNREASONABLE EFFECTIVENESS OF MATHEMATICS IN THE NATURAL SCIENCES, Eugene Wigner, Communications in Pure and Applied Mathematics, Vol. 13, No. 1 (February 1960)

Solve all disease

We're entering a new era of drug discovery — one where frontier AI can unlock deeper scientific insights, faster breakthroughs, and life-changing medicines. At Isomorphic Labs, we're building that future.



TECHNOLOGY > AI

AlphaFold 3 predicts the structure and interactions of all of life's molecules

May 08, 2024

6 min read

Introducing AlphaFold 3, a new AI model developed by Google DeepMind and Isomorphic Labs. By accurately predicting the structure of proteins, DNA, RNA, ligands and more, and how they interact, we hope it will transform our understanding of the biological world and drug discovery.

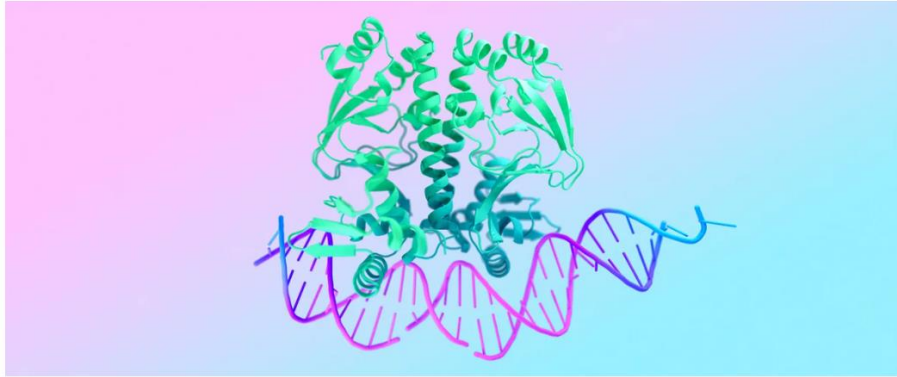


Google DeepMind
AlphaFold team



Isomorphic Labs

Share



nature

Explore content

About the journal

Publish with us

Subscribe

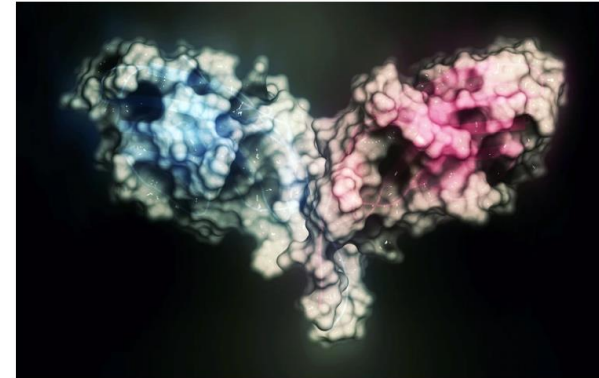
nature > news > article

NEWS | 27 May 2026

Move over, AlphaFold: open-source model predicts shape of 1 billion proteins

An open-source atlas generated by an AI tool called ESMFold2 vastly increases the known protein universe.

By [Ewen Callaway](#) & [Miryam Naddaf](#)



The AI tool designed proteins that would bind to cytotoxic T-lymphocyte-associated protein 4 (CTLA-4).
Credit: Molekuul/SPL

Can machines think? Functional definition → The Imitation Game

A. M. Turing (1950) Computing Machinery and Intelligence. *Mind* 49: 433-460.

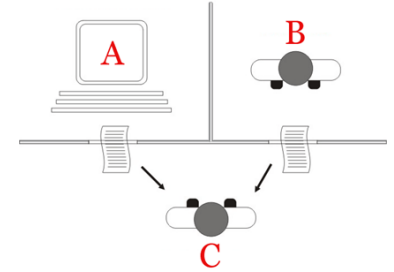
COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

"The game may perhaps be criticised on the ground that the odds are weighted too heavily against the machine. If the man were to try and pretend to be the machine he would clearly make a very poor showing."



Human or Machine moves... that is the question...



Large language models pass a standard three-party Turing test

Cameron R. Jones^{a,b,1}  and Benjamin K. Bergen^a 

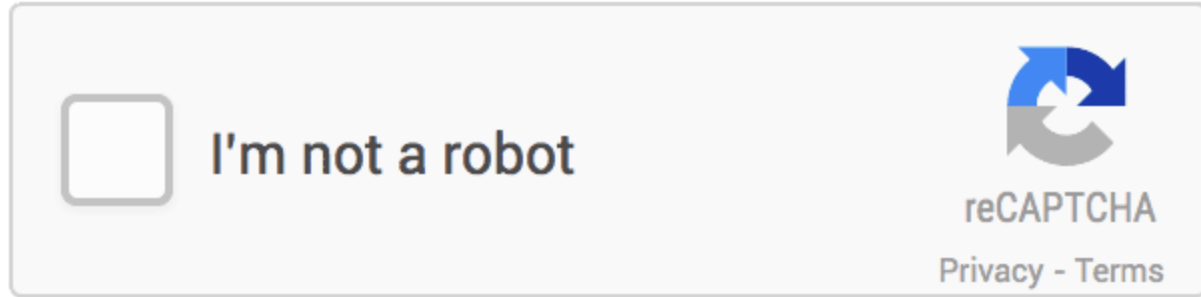
Edited by Brenden Manker Lake, Princeton University, Princeton, NJ; received September 16, 2025; accepted March 27, 2026 by Editorial Board Member Michael S. Gazzaniga

The Turing test has been widely discussed as a test of machine intelligence, but it also provides a measure of how humans distinguish other humans from machines. We evaluated 4 systems (ELIZA, GPT-4o, LLaMa-3.1-405B, and GPT-4.5) in two randomized, controlled, and preregistered Turing tests on independent populations. Participants had 5 min conversations simultaneously with another human participant and one of these systems before judging which conversational partner they thought was human. When prompted to adopt a humanlike persona, GPT-4.5 was judged to be the human 73% of the time: significantly more often than interrogators selected the real human participant. LLaMa-3.1, with the same prompt, was judged to be the human 56% of the time—not significantly more or less often than the humans it was being compared to. Without these prompts, however, the same models performed significantly worse (38% and 36%), and did not consistently outperform baseline models, ELIZA and GPT-4o (23% and 21%, respectively). A third study replicated these results in 15-min games: two PERSONA-prompted models achieved pass rates of 56% and 59%. The results constitute empirical evidence that artificial systems can pass a standard three-party Turing test. Interrogators' reasoning focused more on stylistic and socio-emotional aspects of human behavior rather than more traditional notions of intelligence. The results have implications for debates about what kind of intelligence is exhibited by large language models, the social impacts these systems are likely to have, and the aspects of human behavior that people continue to see as unique.

Significance

The Turing test asks whether a machine can imitate human behavior so well that another human cannot reliably tell the difference. It is not only the oldest and most discussed test of AI but can also provide insight into what cues people use to distinguish humans from machines. This paper demonstrates that—when suitably prompted—three current AI systems achieve a pass rate of at least 50% in a standard Turing test, meaning that participants were no better (and in some cases worse) than chance at selecting between a human and a machine.

Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA)



By Google - <https://www.google.com/recaptcha/intro/index.html>, Fair use, <https://en.wikipedia.org/w/index.php?curid=45398292>

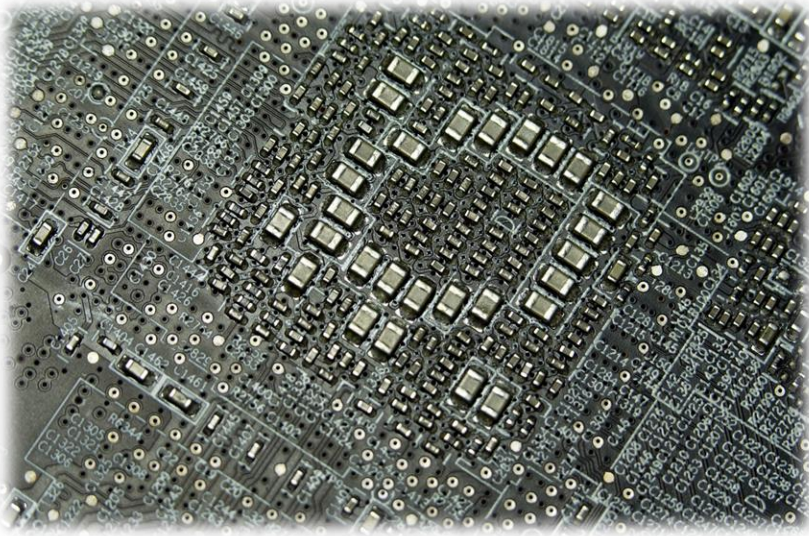
Basic underlying test principles:

- For computational sequential precision tasks in *closed systems* (like chess) the machine outperforms the human
- For representing and processing complex relations in *open systems* (like daily life) the human outperforms the machine

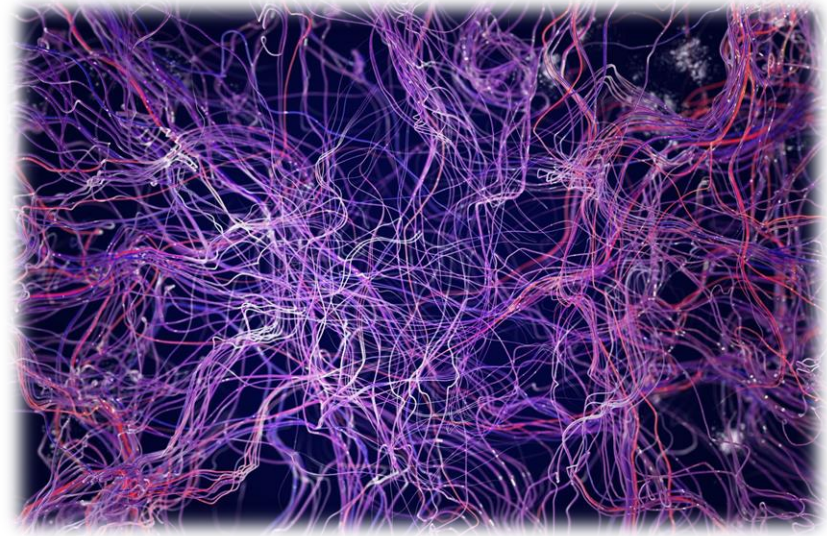
The roles of Machines and Humans

AI vs NI comparison: Silico vs Vivo

Hardware vs Wetware



<https://thewire.in/science/new-transistor-design-combines-best-high-low-performance-versions>

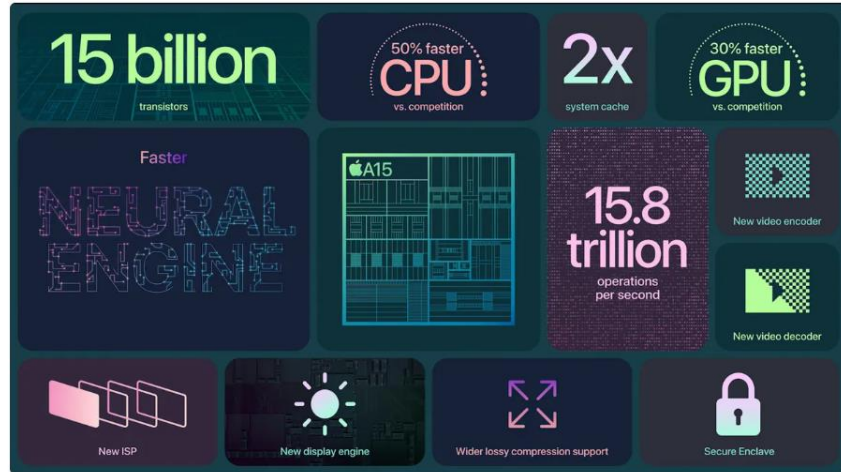


<https://www.ukri.org/news/first-wiring-map-of-neurons-in-insect-brain-complete/>

Machine vs Organism Comparisons

Hardware vs Wetware

Apple's A15 Bionic chip powers iPhone 13 with 15 billion transistors (15×10^9) with 15.8 trillion (15.8×10^{12}) operations per second

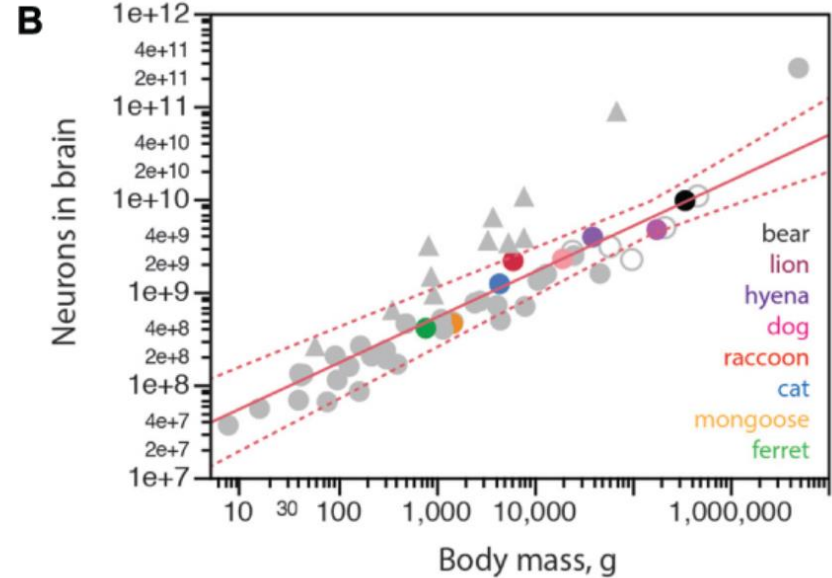


Apple A15 Bionic processor details

Screenshot by Stephen Shankland/CNET

<https://www.cnet.com/tech/mobile/apples-a15-bionic-chip-powers-iphone-13-with-15-billion-transistors-new-graphics-and-ai/>

A cat has around 1.2×10^9 neurons in the brain that individually have a firing frequency between 0.1-300 Hertz, resulting in approximately 10^{10} - 10^{11} operations per second



Front. Neuroanat., 12 December 2017

Volume 11 - 2017 |

<https://doi.org/10.3389/fnana.2017.00118>

Iphone 13 has around 10 times more *compute nodes* and 100-1000 times more *operations per second* than a cat (at maximum capacity)...



The Connectome and parallelism

Science

Current Issue

First release papers

Archive

About ▾

Submit manuscript

HOME > SCIENCE > VOL. 379, NO. 6636 > THE CONNECTOME OF AN INSECT BRAIN

RESEARCH ARTICLE | NEUROSCIENCE



The connectome of an insect brain

MICHAEL WINDING , BENJAMIN D. PEDIGO , CHRISTOPHER L. BARNES , HEATHER G. PATSOLIC , YOUNGSEUNG PARK , AKIRA FUSHIKI , INGRID V. ANDRADE, AVINASH KHANDLWAL , [...], AND MARTA ZLATIC

+10 authors

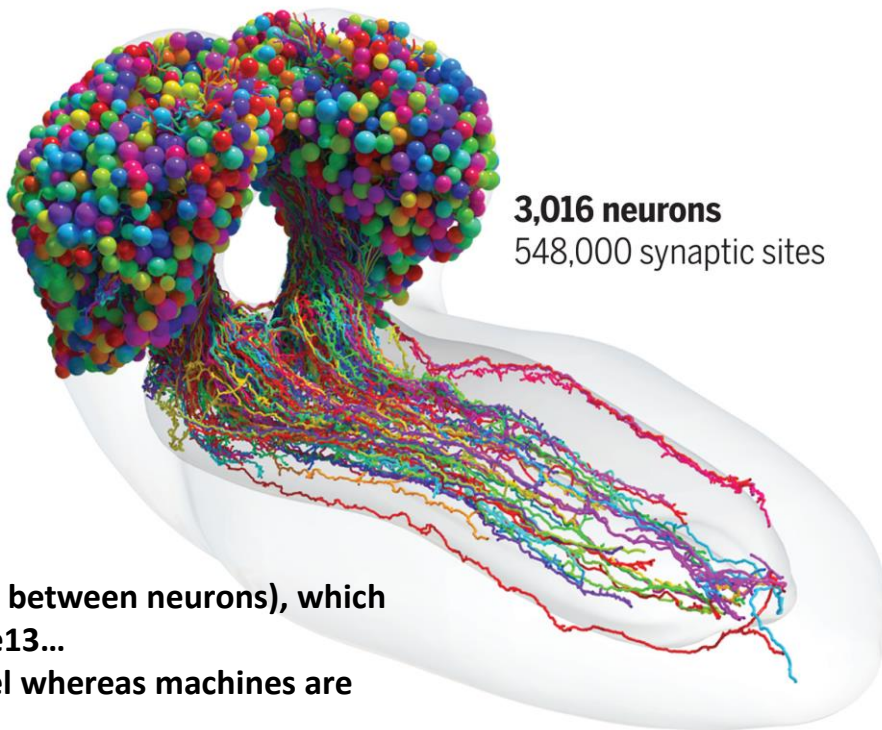
[Auth](#)

[Affiliations](#)

“The complete synaptic-resolution connectome of the Drosophila larval brain comprises 3016 neurons and 548,000 synapses.”

- A cat has approximately 10^{13} synapses (connections between neurons), which is around 200 times more connections than an iPhone13...
- Organism nervous system functioning is highly parallel whereas machines are fundamentally sequential

Morphology



Editorial | Published: 08 December 2025

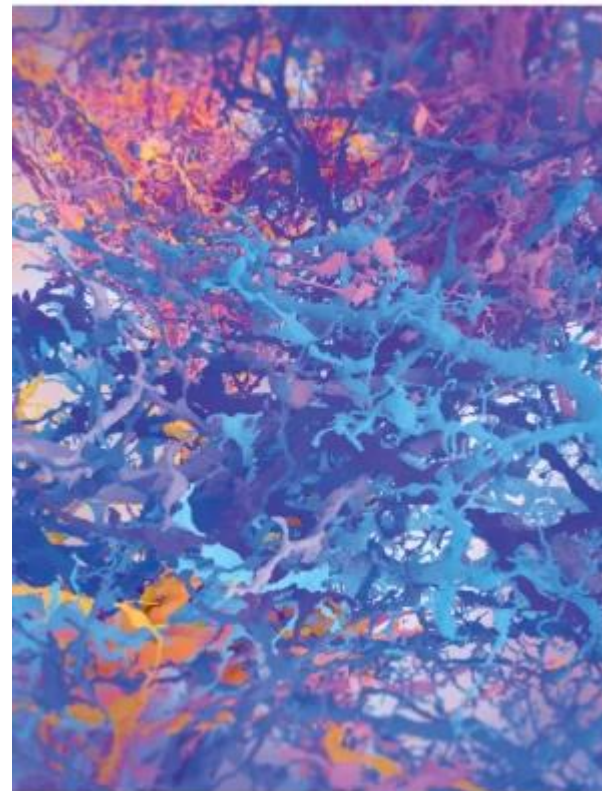
Method of the Year 2025: electron microscopy-based connectomics

[Nature Methods](#) **22**, 2463–2464 (2025) | [Cite this article](#)

25k Accesses | **157** Altmetric | [Metrics](#)

A large network of interconnected neurons serves as the basis of brain function and of behavior. Methodological advances have enabled the reconstruction of large-scale and even whole-brain connectomes of various organisms.

Our ability to think, dream, plan and do critically depends on our brain and its billions of neurons and their trillions of connections. To gain a deeper understanding of the brain, a complete wiring diagram would be desirable, yet given the enormous size of the human brain and the comparatively tiny scale of neurons and synapses, this feat remains elusive. However, impressive strides have been made for smaller brains.



Natural Intelligence -- Animals sleep



[Published: 09 January 2008](#)

Lethargus is a *Caenorhabditis elegans* sleep-like state

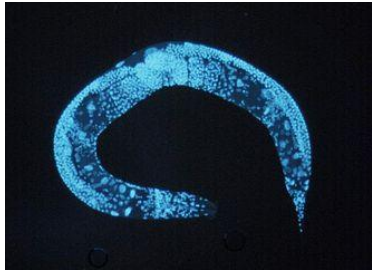
[David M. Raizen](#) , [John E. Zimmerman](#), [Matthew H. Maycock](#), [Uyen D. Ta](#), [Young-jai You](#), [Meera V. Sundaram](#) & [Allan I. Pack](#)

[Nature](#) **451**, 569–572 (2008) | [Cite this article](#)

9534 Accesses | **328** Citations | **80** Altmetric | [Metrics](#)



A [Corrigendum](#) to this article was published on 12 June 2008



Abstract

There are fundamental similarities between sleep in mammals and quiescence in the arthropod *Drosophila melanogaster*, suggesting that sleep-like states are evolutionarily ancient^{1,2,3}. The nematode *Caenorhabditis elegans* also has a quiescent behavioural state during a period called lethargus, which occurs before each of the four moults⁴. Like sleep,

Natural Intelligence -- Animals dream

PNAS

BRIEF REPORT

EVOLUTION

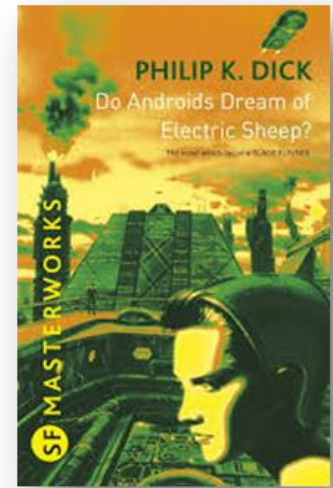
OPEN ACCESS



Regularly occurring bouts of retinal movements suggest an REM sleep-like state in jumping spiders

Daniela C. Rössler^{a,b,c,1}, Kris Kim^d, Massimo De Agrò^e, Alex Jordan^c, C Giovanni Galizia^{a,b}, and Paul S. Shamble^d

Edited by Joan Strassmann, Washington University in St. Louis, St. Louis, MO; received March 22, 2022; accepted May 27, 2022



A jumping spider's legs twitch and curl up when they rest, as seen here, during a REM sleep-like state.

PHOTOGRAPH COURTESY OF DANIELA C. RÖSSLER

“Here, we report evidence for an REM sleep-like state in a terrestrial invertebrate: periodic bouts of retinal movements coupled with limb twitching and stereotyped leg curling behaviors during nocturnal resting in a jumping spider.”

<https://www.nationalgeographic.com/animals/article/jumping-spiders-dream-rem-sleep-study-suggests>

Natural intelligence -- Animals play



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

[Follow this preprint](#)

Seeking voluntary passive movement in flies is play-like behavior

Tilman Triphan, Wolf Huetteroth

doi: <https://doi.org/10.1101/2023.08.03.551880>

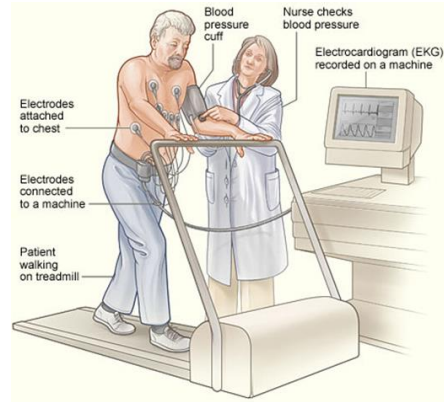
This article is a preprint and has not been certified by peer review [what does this mean?].

Our data show that the flies' active exposure to the carousel fulfils all criteria of play-like behavior (PLB): it is of no immediate relevance for survival, the flies engage with the carousel voluntarily, repeatedly, in a stress-free environment and in a non-ethotypical manner, and the seeking behavior strongly



Drosophila melanogaster

Clinicians have always observed patients and measured

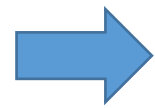
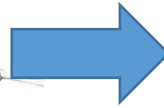


Medicine → Precision Medicine

- Stethoscopes → Microscopes → Deep and holistic biomolecular measurements
- Static measurements of few variables → Longitudinal measurements of many variables
- Human-Machine Partnership for analyses, evaluations and understanding



Game Over → Humans as creative and decision-making artists in the Human-Machine Partnership



- Individual Human Knowledge → Connected Global Knowledge
- Future transformative processes & interfaces
- Legal & Ethical & Psychological implications...

The diffusion of large language models in published academic articles

Kyle Siler   [Authors Info & Affiliations](#)

Edited by Jeffrey D. Ullman, Stanford University, Stanford, CA; received February 19, 2026; accepted April 27, 2026

May 29, 2026 | 123 (22) e2605754123 | <https://doi.org/10.1073/pnas.2605754123>

 2,325



GET ACCESS

Vol. 123 | No. 22

[Significance](#)

Abstract

Data, Materials, and
Software Availability

Acknowledgments

Supporting Information

References

Significance

Understanding how academic knowledge is produced is essential for maintaining public trust in science. Large language models (LLMs) such as ChatGPT are rapidly transforming academic writing, yet little is known about who uses these tools or how extensively. An analysis of 7.3 million journal articles from 2020 to 2025 reveals that by 2025, slightly over half show evidence of LLM influence, with usage varying markedly by world region, institutional prestige, publisher, and discipline. These patterns demonstrate that AI adoption in science is not uniform but shaped by social and institutional forces including language barriers, competitive pressures, and editorial gatekeeping. As AI tools grow more powerful, assessing the provenance and authenticity of academic work will be critical for all stakeholders.



[nature](#) > [nature portfolio](#) > [editorial policies](#) > artificial intelligence (ai)

[Editorial policies](#)

[Acknowledgements](#)

[Act with integrity](#)

[Appeals & Complaints](#)

Artificial Intelligence (AI)

[Authorship](#)

[Clinical Research](#)

[Collections, Guest Edited
Collections, and Special Issues](#)

[Communicate with Respect](#)

[Competing interests](#)

[Confidentiality](#)

[Corrections, Retractions and
Matters Arising](#)

[Funding Statements](#)

[Image integrity and standards](#)

[Peer Review](#)

[Plagiarism and duplicate
publication](#)

[Preprints & Conference
Proceedings](#)

[Press and embargo policies](#)

[Reporting standards and
availability of data, materials, code
and protocols](#)

Artificial Intelligence (AI)

Springer Nature is monitoring ongoing developments in this area closely and will review (and update) these policies as appropriate.

1. [AI authorship](#)
2. [Generative AI images](#)
3. [AI use by peer reviewers](#)
4. [Editorial use](#)

AI authorship

Large Language Models (LLMs), such as ChatGPT, do not currently satisfy our [authorship](#) criteria. Notably an attribution of authorship carries with it accountability for the work, which cannot be effectively applied to LLMs. Use of an LLM should be properly documented in the Methods section (and if a Methods section is not available, in a suitable alternative part) of the manuscript. The use of an LLM (or other AI-tool) for “AI assisted copy editing” purposes does not need to be declared. In this context, we define the term “AI assisted copy editing” as AI-assisted improvements to human-generated texts for readability and style, and to ensure that the texts are free of errors in grammar, spelling, punctuation and tone. These AI-assisted improvements may include wording and formatting changes to the texts, but do not include generative editorial work and autonomous content creation. In all cases, there must be human accountability for the final version of the text and agreement from the authors that the edits reflect their original work.

Modeling and controlling immunodynamics

The Holy grail of modern biomedical research is to *understand and control* the human immune system

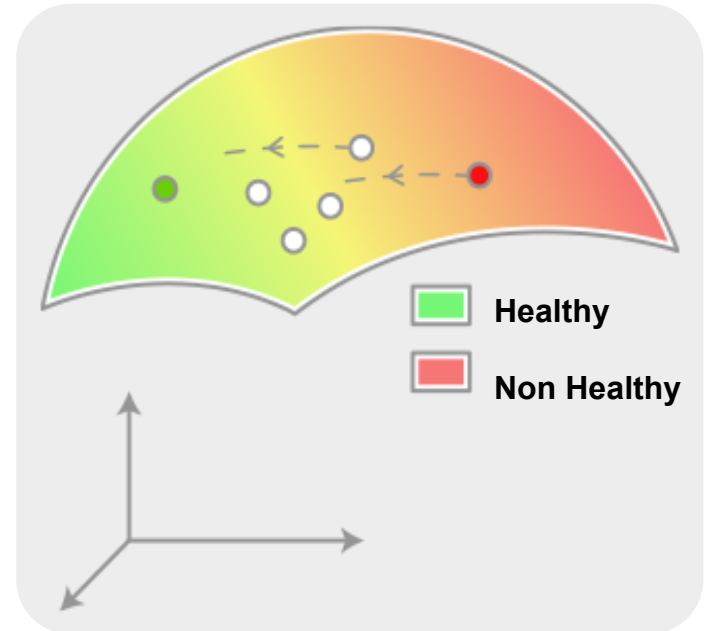


The immune system in

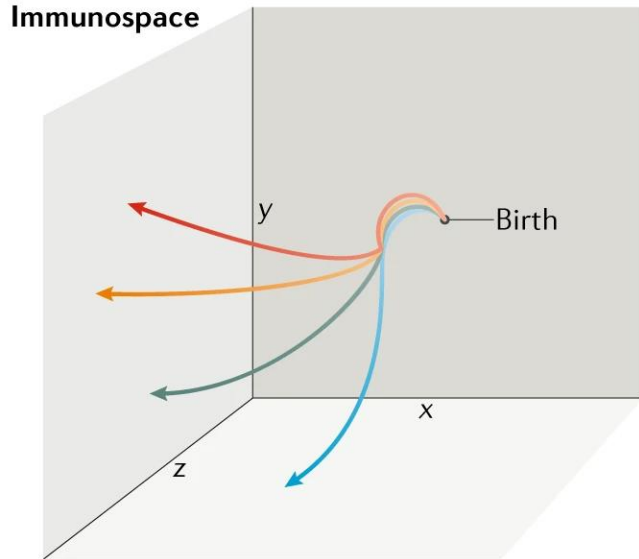
- Autoimmune disease
- Cancer
- Infectious disease
- Neurodegenerative disease
- And most other human disease conditions...

Homeostasis & Perturbations in the Immunity State Space

- *The Immunity State Space* is constructed from « deep/precision longitudinal and integrated measurements » on cells and other biomolecular actors
- Disease (autoimmune, neurodegenerative, cancer) are often connected with **low grade chronic inflammation**, i.e. a new **unfavorable stable attractor** in the state space
- The therapeutic goal is to, through interventions, push patients back to a **favorable (healthy) stable attractor**



General Systems Immunology knowledge suggest e.g. using individuals as their own control



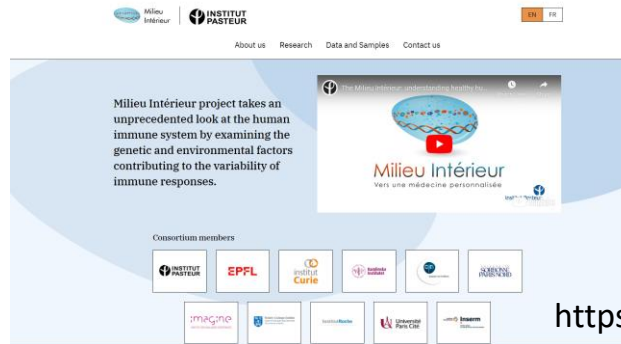
Immune trajectory

- | | |
|--|--|
|  Individual 1 |  Individual 3 |
|  Individual 2 |  Individual 4 |

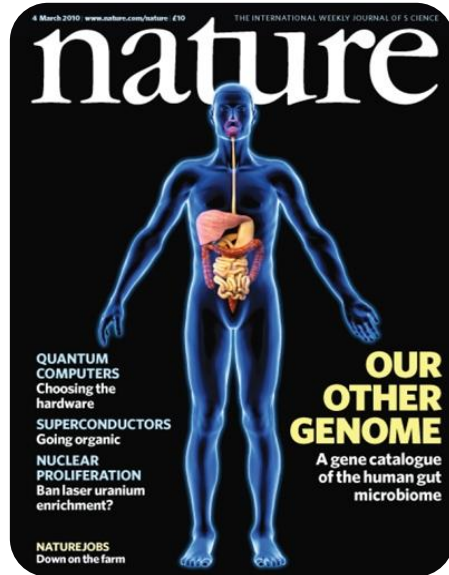
- “Human immune systems are relatively stable within individuals over the course of weeks to months, but incredibly variable between individuals”
- “...induced responses to pathogens differ markedly among different age groups and... these differences are unique to different kinds of stimuli”

Petter Brodin

Nature Reviews | Immunology volume 19 | February 2019 | 87



The human model is complex and context dependent



Nature March 2010

Nature June 2012

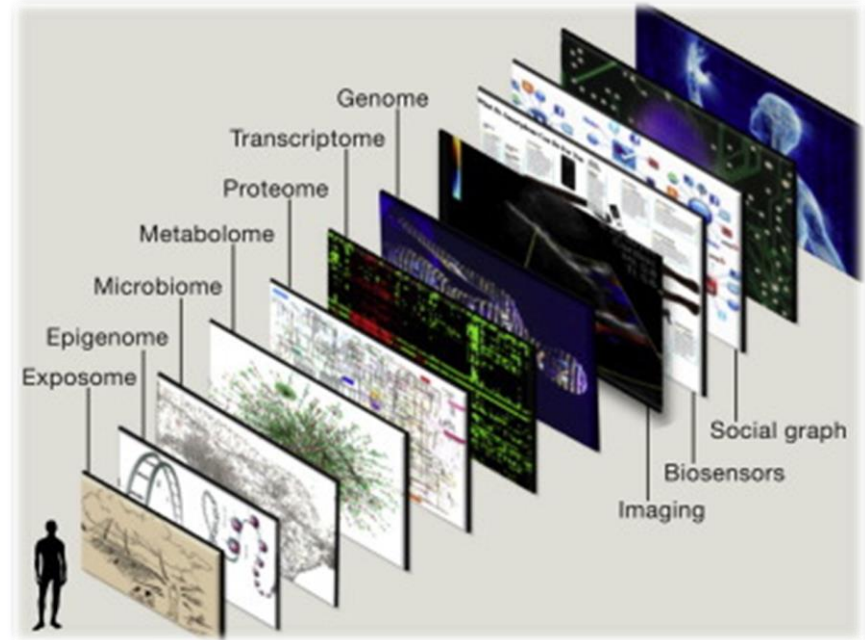
- 3×10^{13} human cells
- As many bacterial cells
- Around 10 times as many viruses.

A very dynamical system...

The complex « *Human Dynamical System* »

Human biology might be modeled as a hierarchical system of connected and interacting *dynamical complex systems* of

- ...
- Molecules $\leftrightarrow$
- Cells $\leftrightarrow$
- Tissues $\leftrightarrow$
- Organs $\leftrightarrow$
- Organ complexes $\leftrightarrow$
- Organisms $\leftrightarrow$
- Populations $\leftrightarrow$
-



40⁺ Leading Edge
Review

Individualized Medicine
from Prewomb to Tomb

Eric J. Topol^{1,2}
¹The Scripps Translational Science Institute, The Scripps Research Institute and Scripps Health, La Jolla, CA 92037, USA
²Correspondence: etopol@scripps.edu
https://doi.org/10.1016/j.elsevier.2014.02.012

Cell

Time dynamics: ~ 4 million human cells die (and are created) per second in a human body

nature
medicine

BRIEF COMMUNICATION

<https://doi.org/10.1038/s41591-020-01182-9>



The distribution of cellular turnover in the human body

Ron Sender and Ron Milo

We integrated ubiquity, mass and lifespan of all major cell types to achieve a comprehensive quantitative description of cellular turnover. We found a total cellular mass turnover of 80 ± 20 grams per day, dominated by blood cells and gut epithelial cells. In terms of cell numbers, close to 90% of the $(0.33 \pm 0.02) \times 10^{12}$ cells per day turnover was blood cells.

To better understand the function of the human body in health and disease, it is of major interest to quantify its cellular compo-

cells comprising the human body¹ or ones with an especially fast turnover of $\tau < 10$ d.

We analyzed many of the tissues thought to be relevant and found them to make a negligible contribution in terms of both number and mass (Supplementary Tables 1–4); for example, sperm cells, kidney cells and osteocytes. For cell types with a short lifespan, we revised earlier estimates^{1,3} of the total cell number, as documented in Methods. Figure 1a presents for each of these cell types the number

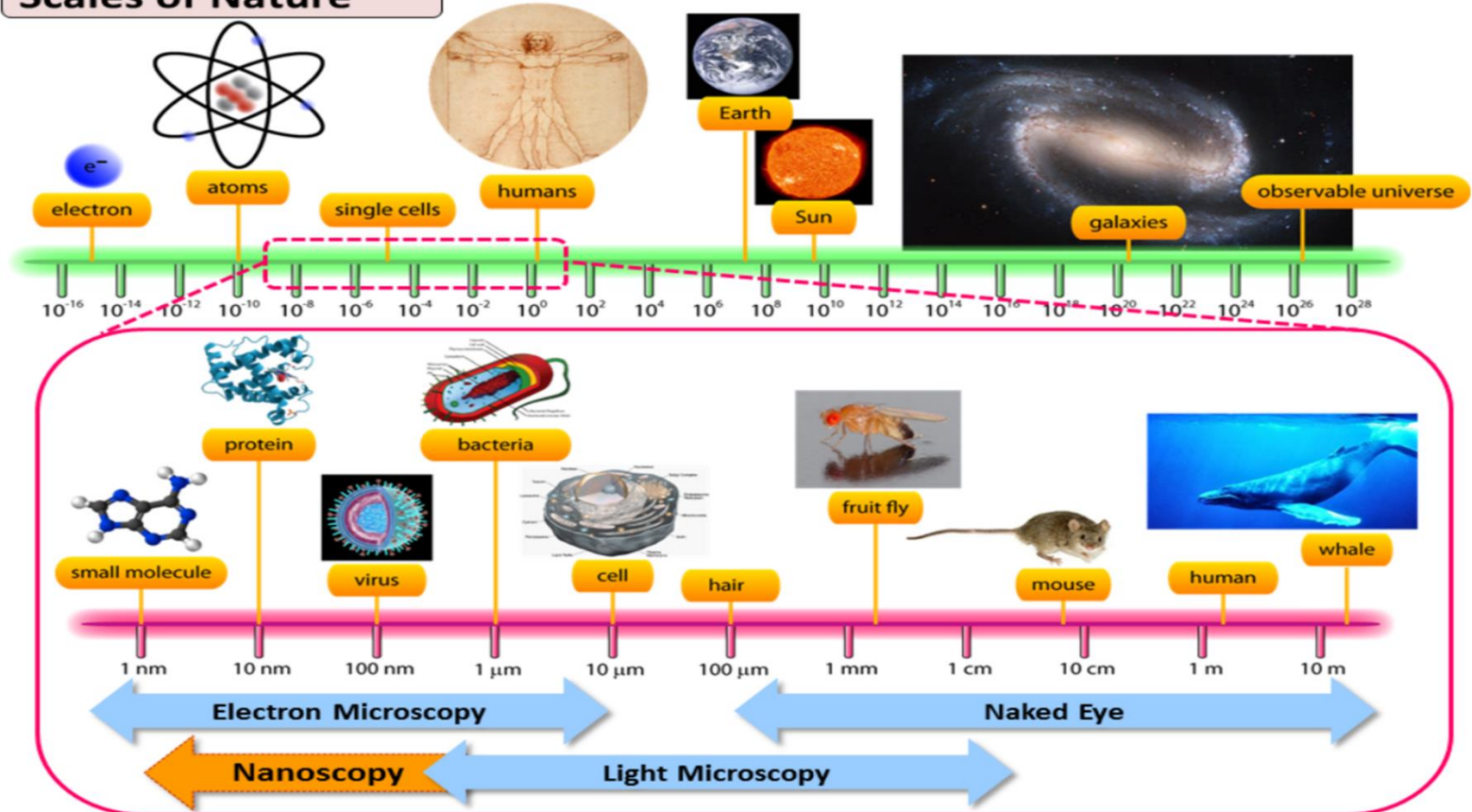
“We found a total cellular mass turnover of 80 ± 20 grams per day, dominated by blood cells and gut epithelial cells. In terms of cell numbers, close to 90% of the $(0.33 \pm 0.02) \times 10^{12}$ cells per day turnover was blood cells.”

Nature Medicine | VOL 27 | January 2021 | 45–48 |
www.nature.com/naturemedicine

“The most recent rigorous census has estimated the number of cells in the adult body to be approximately $30 \pm 0.5 \times 10^{12}$, of which about 90% are from the hematopoietic lineage, mostly red blood cells.”

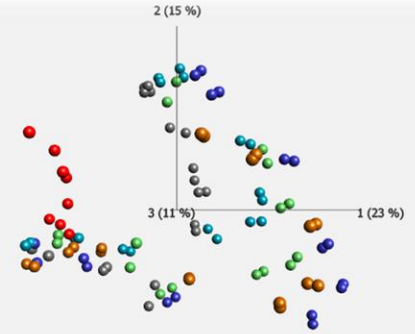
Some (log) length scales of Life

Scales of Nature



Computational modeling experience tells us that « *Life operates in low dimension* »

A typical example: Invitro stimulated mouse dendritic cells 0-0.5-2-4-6-8-12-16- 24h bulk mRNA expression for all 10716 genes (two samples per time point) result in approximately 1 dimensional trajectories embedded in low (<7) dimensional space:



Science

NEWS CAREERS COMMENTARY JOURNALS

Science

Current Issue First release papers Archive About

Submit manuscript

HOME > SCIENCE > VOL. 326, NO. 5950 > UNBIASED RECONSTRUCTION OF A MAMMALIAN TRANSCRIPTIONAL NETWORK MEDIATING PATHOGEN RESPONSES

RESEARCH ARTICLE

Unbiased Reconstruction of a Mammalian Transcriptional Network Mediating Pathogen Responses

IDO AMIT, MANUEL GARBER, NICOLAS CHEVRIER, ANA PAULA LEITE, YONI DONNER, THOMAS EISENHAURE, MITCHELL GUTTMAN, JENNIFER K. GRENIER, WEIBO LI, [...], AND AVIV REGEV +15 authors [Authors Info & Affiliations](#)

SCIENCE • 3 Sep 2009 • Vol 326, Issue 5950 • pp. 257-263 • DOI: 10.1126/science.1179050

NCBI GEO

COVID-19 is an emerging, rapidly evolving situation. Get the latest public health information from CDC: <https://www.cdc.gov/covid>. Get the latest research from NIH: <https://www.nih.gov/covid19>. Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>

NCBI - GEO - Accession Display | GEO Publications - FAQ - HOME - Email GEO

NCBI help: Mouse over screen elements for information. not logged in | Login

Scope: (Self) Format: (HTML) Amount: (Quick) GEO accession: GSE117721

Series GSE117721 Query Datasets for GSE117721

Status Public on Oct 09, 2009
Title Dendritic cell stimulation: 5 ligands, 9 time points
Organism Mus musculus
Experiment type Expression profiling by array
Summary Mouse primary BMDMs were stimulated with TLR ligands and gene expression changes were profiled on Affymetrix arrays
Overall design BMDMs were stimulated with 5 TLR ligands (LPS, pIC, PAM, CpG, GRD) across 9 time points (0, 5, 1, 2, 4, 6, 8, 12, 16, 24 hours). Unstimulated cells were used as controls.
Contributor(s) Amit I
Citation(s) Amit I, Garber M, Chevrier N, Leite AP et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science 2009 Oct 9;326(5950):257-63. PMID: 19729616
Rapp S, Barnett FW, Barnett-Dahms Z, Amit I et al. Analysis of the transcriptional networks underlying the activation of murine macrophages by inflammatory mediators. J Leukoc Biol 2014 Aug;96(2):167-83. PMID: 24771704

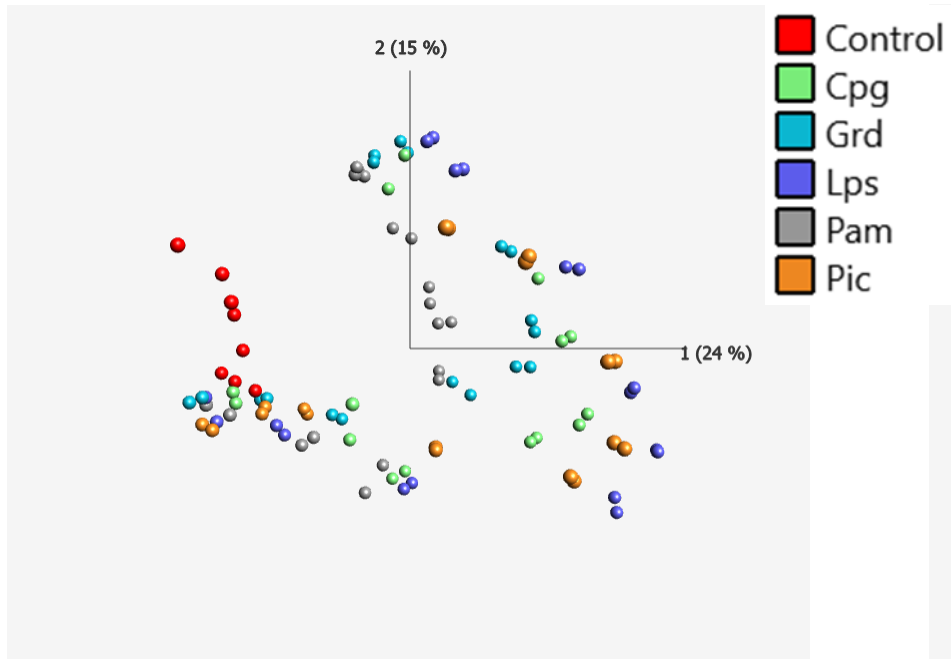
- Control
- Cpg
- Grd
- Lps
- Pam
- Pic

A small sub-part of the full 96 times 10716 matrix, representing 10716 genes and 96 samples for the invitro stimulated mouse dendritic cell data:

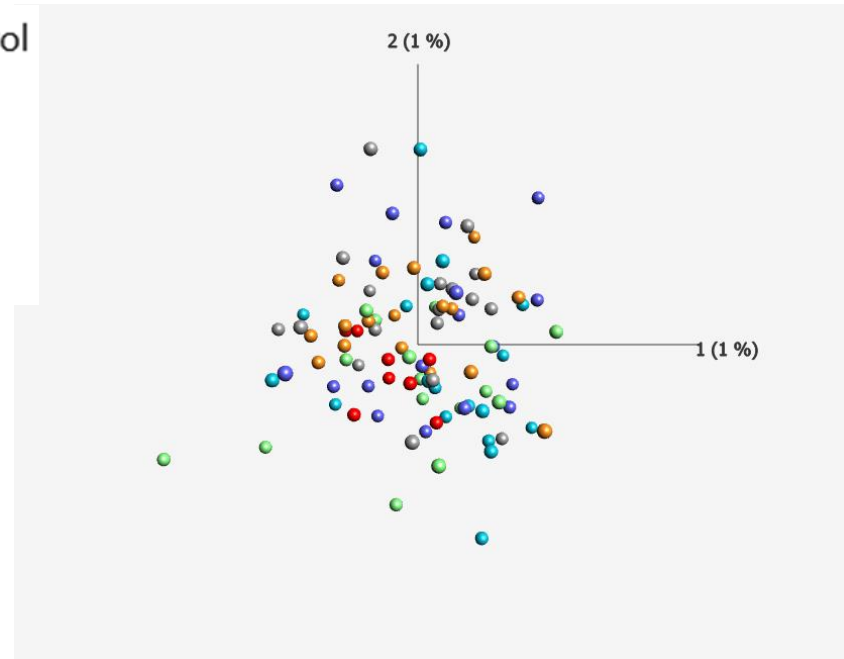
	p-values	q-values	ol	Control	Control	Control	Control	Control	Control	Control	Cpg	Cpg
Myo5a				374.6	381.35	408.6	387.6	424.55	411.5	396.1	376.25	360.65
Nabp1			5	685.6	711.7	659.55	718.5	758.75	880.05	758.95	764.2	761.05
Nacc2				66.4	36.7	44.5	43.6	39.5	41.8	44.3	38.6	40
Myo1c				157.7	189.85	133.9	178.45	199.55	142.25	174.95	190.35	165.45
Myof				726.6	635.1	610.1	615.45	639.2	624.8	606.55	639.45	629.5
Mynn				31.35	34.05	29.8	28.3	31.1	35.1	36.6	27.55	27.05
Naca			4	2856.3	2685.5	3526.1	2863.4	2820.9	2867.7	2814.1	2989.4	2926.1
Naa35				71.8	61.9	78.9	70.4	70.6	64.4	73.5	66.4	66.6
Nab1			5	366.95	339.65	444.75	331.15	334.9	372	342.05	300.55	292.1
Mycl				20	20	20	20	20	20	20	20	20
Ndfip1				1162.4	1003.4	1124.4	932.4	1098.5	1105.8	1229.8	957.5	941.5
Nap111///...				348.35	525.35	401.3	539.1	445.1	389.75	416.65	612	550.55
Ncor1				557.9	409.9	548.9	447.8	437.3	460.9	414.8	435.9	425
Ncaph2				142.4	140.5	139.8	124.5	134.7	153.5	151.8	142.8	155.6
Ndel1			5	305.55	336.45	297.55	320.25	298.05	276.65	306.2	328.9	322.9

How to analyze and visualize high dimensional measurements?

Illustrating that life operates in low dimension...



PCA on the real data set capturing 39% of the variance in 2D (50% in 3D)



PCA on corresponding randomized data capturing 2% of the variance in 2D (3% in 3D)

**STATISTICAL LEARNING & VISUALIZATION →
ACQUIRING INSIGHTS AROUND PATTERNS IN DATA
RELATED TO BIOLOGICAL VARIATION**

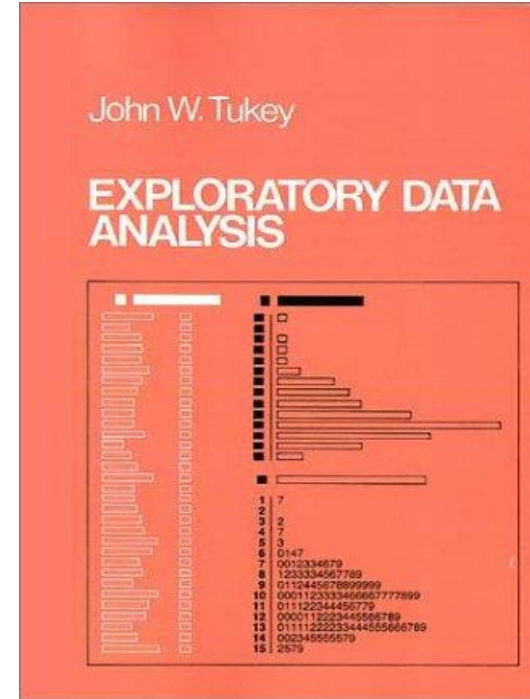


Design and Photo: Emilia Fontes 2016

Exploratory Data Analysis vs Confirmatory Data Analysis



John Wilder Tukey (1915-2000)
Inventor of the FFT, the Box plot and the word
“bit”.



1977

First question: What should it mean to be similar?

-> Choice of **similarity measure or distance function**

Example genuine metrics:

A **metric space** is a set M together with a fixed **distance function** or **metric** $d : M \times M \longrightarrow [0, \infty)$ such that for all x, y and z in M we have

$$d(x, y) \geq 0 \quad \text{with equality if and only if (iff) } x = y \quad (2.1)$$

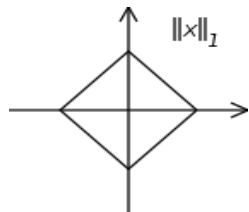
$$d(x, y) = d(y, x) \quad (\textit{symmetry}) \quad (2.2)$$

$$d(x, y) \leq d(x, z) + d(z, y) \quad (\textit{the triangle inequality}) \quad (2.3)$$

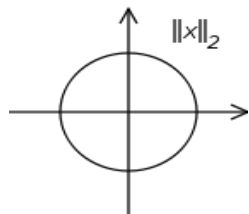
Examples:

- **Different types of edit distances** for e.g. sequencing data
- **L2** or *Euclidean distance* for quantitative data
- **L1** or *Manhattan (or Taxicab) distance*

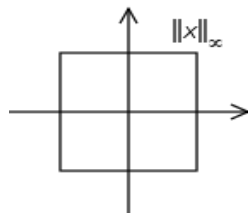
Effects of different similarity measures or distance functions



The “unit sphere” in the L^1 norm. A constraint formulated with this norm favors *sparse vectors*. The idea behind the *Lasso*, see Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288).



The normal “unit sphere” in the standard L^2 norm



The “unit sphere” in the L^∞ norm

[illegible]

A collection of 20 small, colorful, pixelated figures arranged in a grid-like pattern on a light-colored wooden surface. The figures are designed to look like various Pokémon, including Pikachu, Eevee, and others, rendered in a low-resolution, blocky style.



How to explore MULTI-OMICS data?

[illegible]

- **N=10-10⁵ Samples**
- **P=10-10¹⁰ Variables**

How to find relevant structure?

Problems:

- **Noise**
- **Artifacts: technical, batch effects,..**
- **High Dimension**
- **Often few samples and many variables**

Visualize and Analyze the data

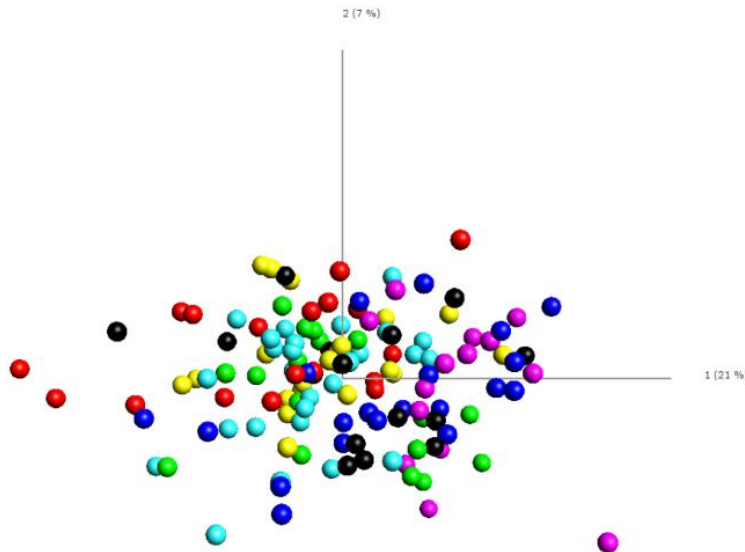
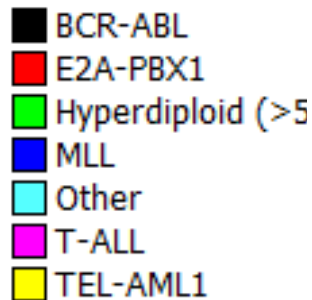
- Choice of “*similarity measure*”
- Choice of dimension/model-reduction method = choice of “*objective function*”

Mary E. Ross et al.

Data available at <https://www.stjuderesearch.org/site/data/ALL3/>

	nsolvent	nsolvent	ICR-41	ICR-42	ICR-43	ICR-44	ICR-45	ICR-46	ICR-47	ICR-48	ICR-49	ICR-50	ICR-51	ICR-52	ICR-53	ICR-54	ICR-55	ICR-56	ICR-57	ICR-58	ICR-59	ICR-60	ICR-61	ICR-62	ICR-63	ICR-64	ICR-65	ICR-66	ICR-67	ICR-68	ICR-69	ICR-70	ICR-71	ICR-72	ICR-73	ICR-74	ICR-75	ICR-76	ICR-77	ICR-78	ICR-79	ICR-80	ICR-81	ICR-82	ICR-83	ICR-84	ICR-85	ICR-86	ICR-87	ICR-88	ICR-89	ICR-90	ICR-91	ICR-92	ICR-93	ICR-94	ICR-95	ICR-96	ICR-97	ICR-98	ICR-99	ICR-100	ICR-101	ICR-102	ICR-103	ICR-104	ICR-105	ICR-106	ICR-107	ICR-108	ICR-109	ICR-110	ICR-111	ICR-112	ICR-113	ICR-114	ICR-115	ICR-116	ICR-117	ICR-118	ICR-119	ICR-120	ICR-121	ICR-122	ICR-123	ICR-124	ICR-125	ICR-126	ICR-127	ICR-128	ICR-129	ICR-130	ICR-131	ICR-132	ICR-133	ICR-134	ICR-135	ICR-136	ICR-137	ICR-138	ICR-139	ICR-140	ICR-141	ICR-142	ICR-143	ICR-144	ICR-145	ICR-146	ICR-147	ICR-148	ICR-149	ICR-150	ICR-151	ICR-152	ICR-153	ICR-154	ICR-155	ICR-156	ICR-157	ICR-158	ICR-159	ICR-160	ICR-161	ICR-162	ICR-163	ICR-164	ICR-165	ICR-166	ICR-167	ICR-168	ICR-169	ICR-170	ICR-171	ICR-172	ICR-173	ICR-174	ICR-175	ICR-176	ICR-177	ICR-178	ICR-179	ICR-180	ICR-181	ICR-182	ICR-183	ICR-184	ICR-185	ICR-186	ICR-187	ICR-188	ICR-189	ICR-190	ICR-191	ICR-192	ICR-193	ICR-194	ICR-195	ICR-196	ICR-197	ICR-198	ICR-199	ICR-200	ICR-201	ICR-202	ICR-203	ICR-204	ICR-205	ICR-206	ICR-207	ICR-208	ICR-209	ICR-210	ICR-211	ICR-212	ICR-213	ICR-214	ICR-215	ICR-216	ICR-217	ICR-218	ICR-219	ICR-220	ICR-221	ICR-222	ICR-223	ICR-224	ICR-225	ICR-226	ICR-227	ICR-228	ICR-229	ICR-230	ICR-231	ICR-232	ICR-233	ICR-234	ICR-235	ICR-236	ICR-237	ICR-238	ICR-239	ICR-240	ICR-241	ICR-242	ICR-243	ICR-244	ICR-245	ICR-246	ICR-247	ICR-248	ICR-249	ICR-250	ICR-251	ICR-252	ICR-253	ICR-254	ICR-255	ICR-256	ICR-257	ICR-258	ICR-259	ICR-260	ICR-261	ICR-262	ICR-263	ICR-264	ICR-265	ICR-266	ICR-267	ICR-268	ICR-269	ICR-270	ICR-271	ICR-272	ICR-273	ICR-274	ICR-275	ICR-276	ICR-277	ICR-278	ICR-279	ICR-280	ICR-281	ICR-282	ICR-283	ICR-284	ICR-285	ICR-286	ICR-287	ICR-288	ICR-289	ICR-290	ICR-291	ICR-292	ICR-293	ICR-294	ICR-295	ICR-296	ICR-297	ICR-298	ICR-299	ICR-300	ICR-301	ICR-302	ICR-303	ICR-304	ICR-305	ICR-306	ICR-307	ICR-308	ICR-309	ICR-310	ICR-311	ICR-312	ICR-313	ICR-314	ICR-315	ICR-316	ICR-317	ICR-318	ICR-319	ICR-320	ICR-321	ICR-322	ICR-323	ICR-324	ICR-325	ICR-326	ICR-327	ICR-328	ICR-329	ICR-330	ICR-331	ICR-332	ICR-333	ICR-334	ICR-335	ICR-336	ICR-337	ICR-338	ICR-339	ICR-340	ICR-341	ICR-342	ICR-343	ICR-344	ICR-345	ICR-346	ICR-347	ICR-348	ICR-349	ICR-350	ICR-351	ICR-352	ICR-353	ICR-354	ICR-355	ICR-356	ICR-357	ICR-358	ICR-359	ICR-360	ICR-361	ICR-362	ICR-363	ICR-364	ICR-365	ICR-366	ICR-367	ICR-368	ICR-369	ICR-370	ICR-371	ICR-372	ICR-373	ICR-374	ICR-375	ICR-376	ICR-377	ICR-378	ICR-379	ICR-380	ICR-381	ICR-382	ICR-383	ICR-384	ICR-385	ICR-386	ICR-387	ICR-388	ICR-389	ICR-390	ICR-391	ICR-392	ICR-393	ICR-394	ICR-395	ICR-396	ICR-397	ICR-398	ICR-399	ICR-400	ICR-401	ICR-402	ICR-403	ICR-404	ICR-405	ICR-406	ICR-407	ICR-408	ICR-409	ICR-410	ICR-411	ICR-412	ICR-413	ICR-414	ICR-415</
--	----------	----------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	-----------

Dimension reduction techniques: *Classical PCA*



ALL (Leukemia) samples
mRNA bulk

22282 variables 132 samples

**PCA based on the
correlation
matrix;**

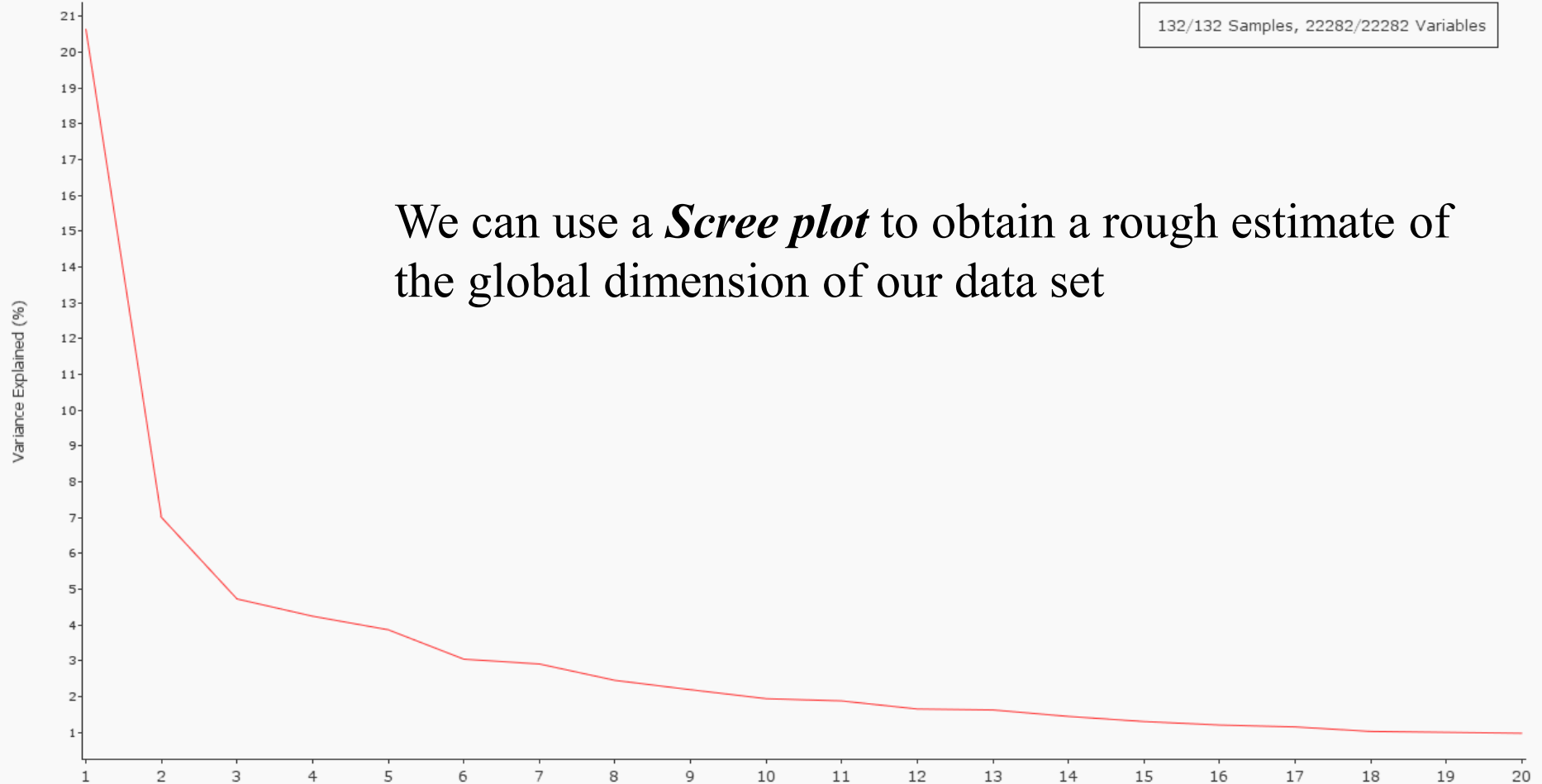
Mean centering
and normalizing all
variables to unit
variance

Distances: Euclidean

Objective function: Total Variance

Scree Plot

132/132 Samples, 22282/22282 Variables



RESEARCH ARTICLE

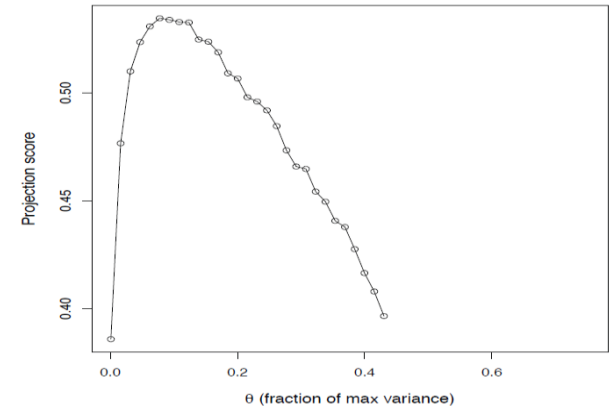
Open Access

The projection score - an evaluation criterion for variable subset selection in PCA visualization

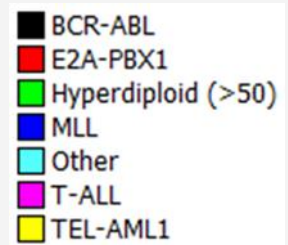
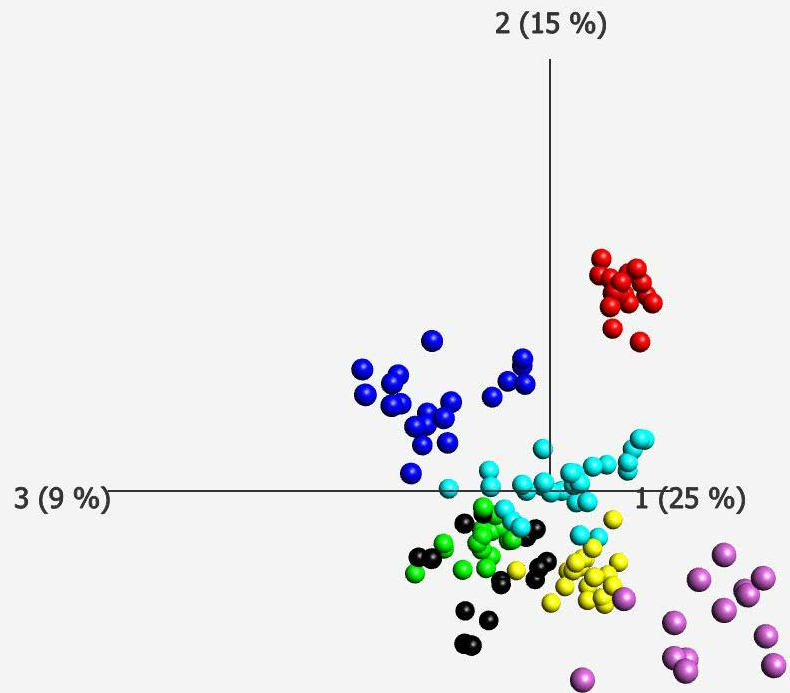
Magnus Fontes* and Charlotte Soneson

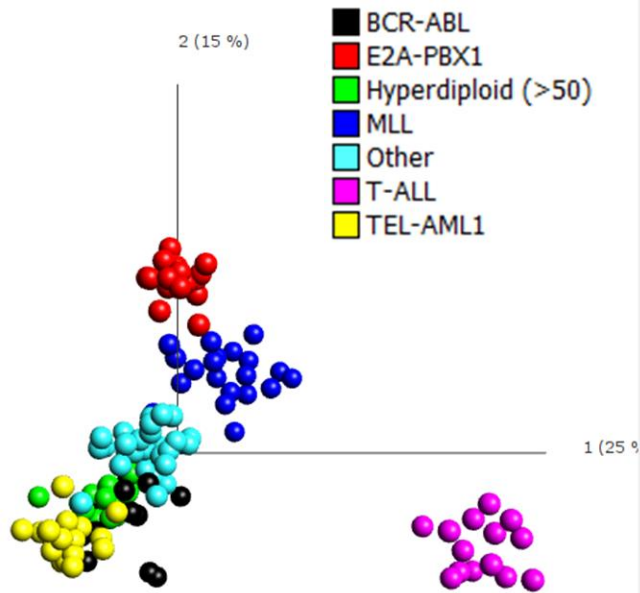
$$\alpha_2(\mathbf{\Lambda}_X, S) = \frac{\sum_{k \in S} \lambda_k^2}{\sum_{k=1}^r \lambda_k^2}.$$

$$\tau(\phi_m(\mathbf{X}), S, \mathcal{P}_{\phi_m(\mathbf{X})}) = (\alpha_2(\mathbf{\Lambda}_{\phi_m(\mathbf{X})}, S))^{1/2} \\ - \mathbb{E}_{\mathcal{P}_{\phi_m(\mathbf{X})}} \left[(\alpha_2(\mathbf{\Lambda}_{\phi_m(\mathbf{X})}, S))^{1/2} \right].$$

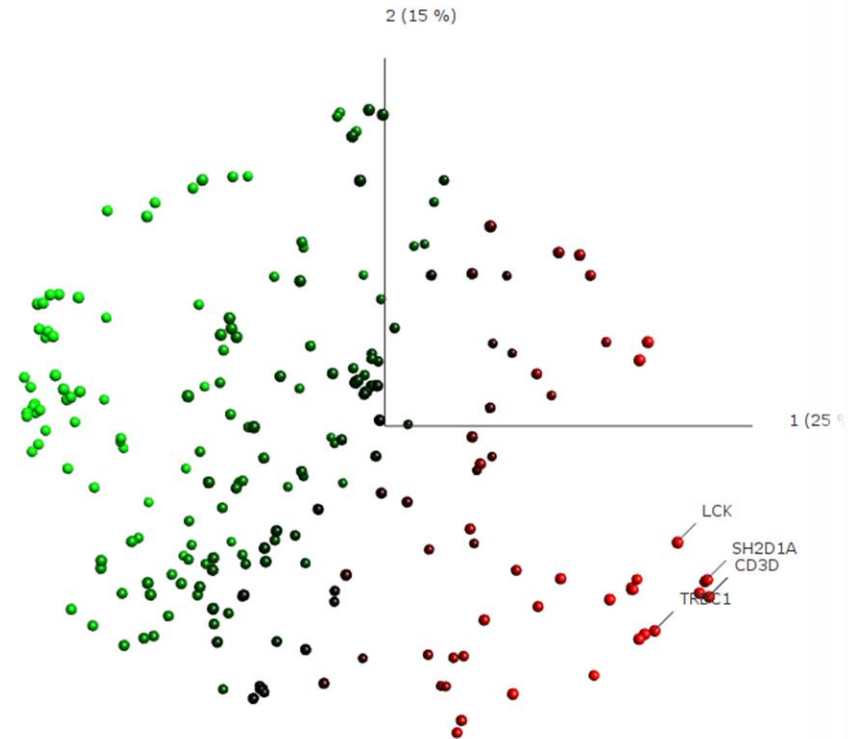


Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO: Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000, **24**:227-235.



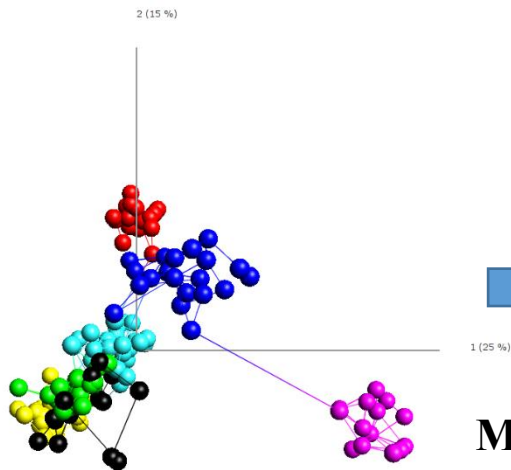


Duality of PCA $\rightarrow$ *Bi-plot* after optimization over *projection score*



A Global Geometric Framework for Nonlinear Dimensionality Reduction

Joshua B. Tenenbaum,^{1*} Vin de Silva,² John C. Langford³



■ BCR-ABL
■ E2A-PBX1
■ Hyperdiploid (>50)
■ MLL
■ Other
■ T-ALL
■ TEL-AML1

MDS based on graph distances

ISOMAP

isomap

Isometric Feature Mapping Ordination

The function performs isometric feature mapping which consists of three simple steps: (1) retain only some of the shortest dissimilarities among objects; (2) estimate dissimilarities as shortest path distances; and (3) perform metric scaling (Tenenbaum et al. 2000).

Keywords [multivariate](#)

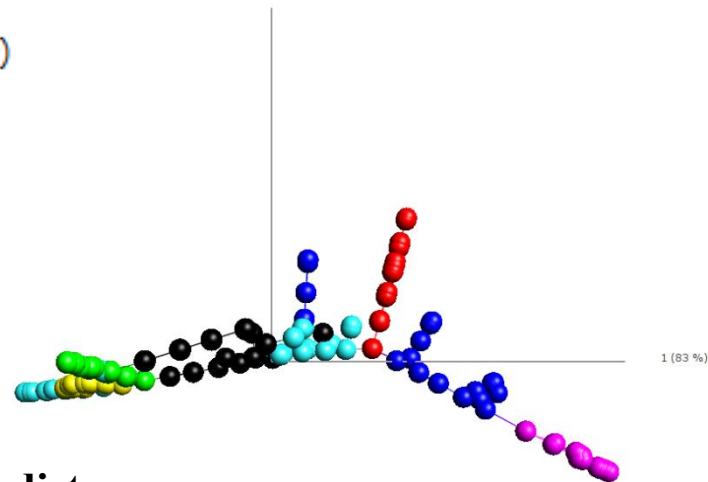
Usage

```
isomap(dist, ndim=2, ...)  
isomap(dist, epsilon, k, path = "shortest", fragmentsize = NULL, ...)  
summary(object, axes = 4, ...)  
plot(x, rot = TRUE, col = "gray", type = "points", ...)
```

Arguments

dist Dissimilarities.
ndim Number of axes in metric scaling (argument [k](#) in [isomap](#)).
epsilon Shortest dissimilarity retained.

2 (8 %)



Stochastic Neighbor Embedding (SNE)

Original distances

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)},$$

Original SNE

Sam Roweis and Geoffrey Hinton

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}.$$

Distances in reduced space

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}.$$

tSNE

Van der Maaten

The cost function C is given by

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

Kullback Leibler
Relative Entropy

Package 'Rtsne'

June 30, 2016

Type Package

Title T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation

Version 0.11

Description An R wrapper around the fast T-distributed Stochastic Neighbor Embedding implementation by Van der Maaten.

License BSD_3_clause + file LICENSE

URL <https://github.com/jkrijthe/Rtsne>

Imports Rcpp (>= 0.11.0)

LinkingTo Rcpp

Suggests testthat

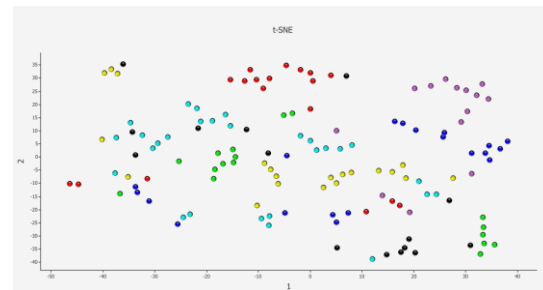
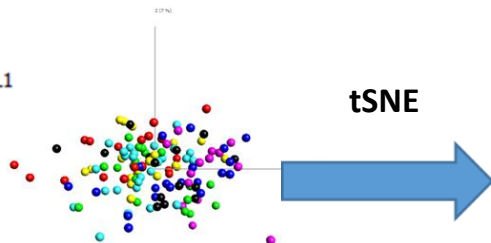
RoxyenNote 5.0.1

NeedsCompilation yes

Author Jesse Krijthe [aut, cre],
Laurens van der Maaten [cph] (Author of original C++ code)

Stochastic Neighbor Embedding (SNE) and tSNE

■ BCR-ABL
 ■ E2A-PBX1
 ■ Hyperdiploid (>50)
 ■ MLL
 ■ Other
 ■ T-ALL
 ■ TEL-AML1



Journal of Machine Learning Research 1 (2008) 1-48

Submitted 4/00; Published 1/01

Visualizing Data using t-SNE

Laurens van der Maaten

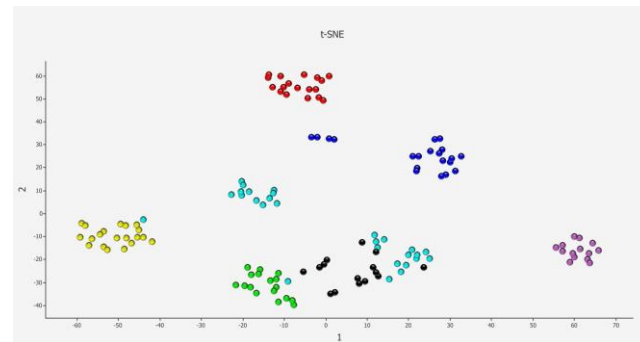
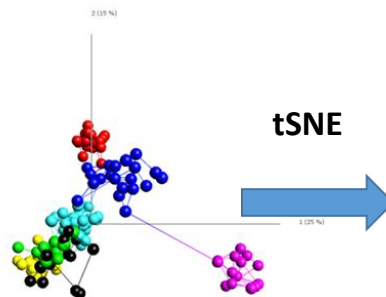
MICC-IKAT
Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands

Geoffrey Hinton

Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

L.VANDERMAATEN@MICC.UNIMAAS

HINTON@CS.TORONTO.E





latest

USER GUIDE / TUTORIAL:

- How to Use UMAP
- Basic UMAP Parameters
- Plotting UMAP results
- UMAP Reproducibility
- Transforming New Data with UMAP
- Inverse transforms
- UMAP on sparse data
- UMAP for Supervised Dimension Reduction and Metric Learning
- Using UMAP for Clustering
- Outlier detection using UMAP
- Document embedding using UMAP
- Embedding to non-Euclidean spaces

Read the Docs

v: latest

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique can be used for visualisation similarly to t-SNE, but also for general non-linear dimension reduction. The algorithm is founded on three assumptions about the data

- The data is uniformly distributed on Riemannian manifold;
- The Riemannian metric is locally constant (or can be approximated as such);
- The manifold is locally connected.

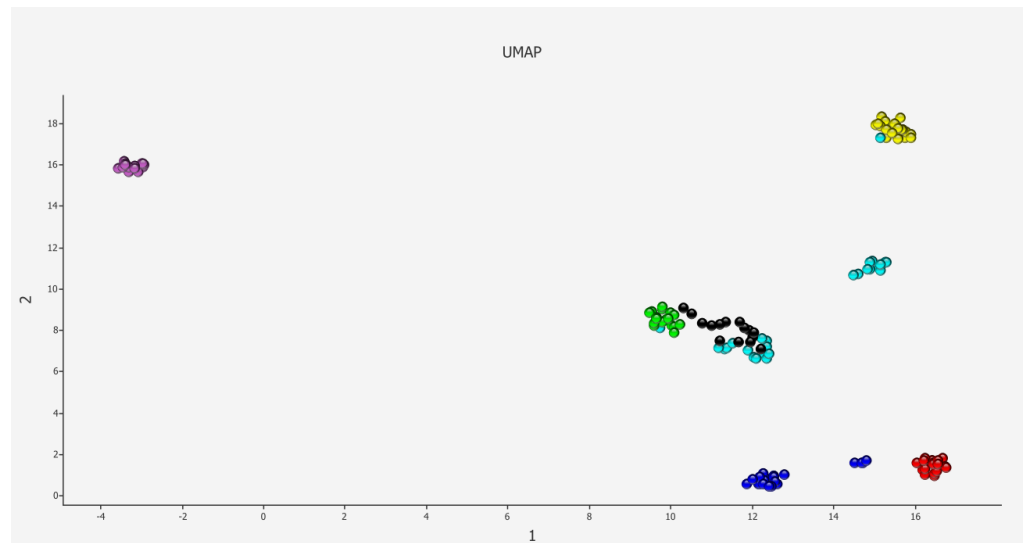
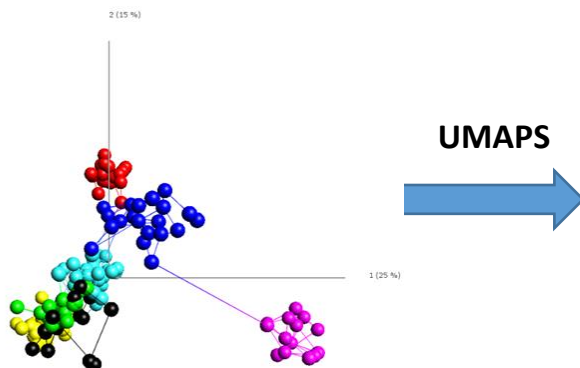
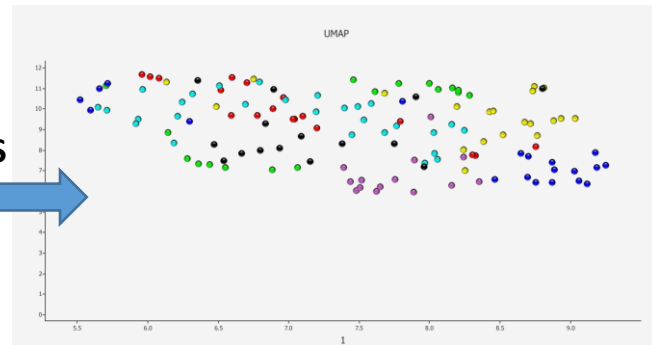
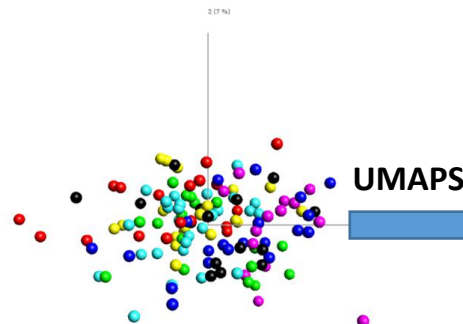
From these assumptions it is possible to model the manifold with a fuzzy topological structure. Embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure.

The details for the underlying mathematics can be found in our [paper on ArXiv](#):

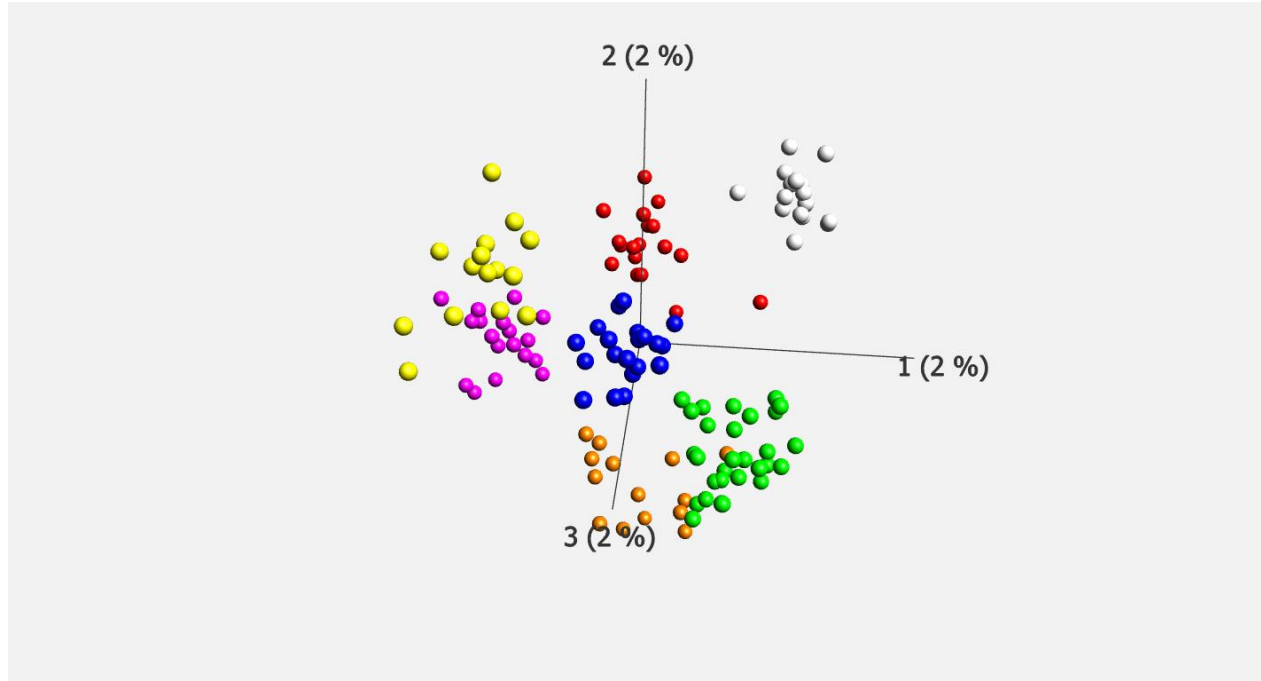
McInnes, L. Healy, J. UMAP: *Uniform Manifold Approximation and Projection for Dimension Reduction* e-prints 1802.03426, 2018

You can find the software [on github](#).

Installation



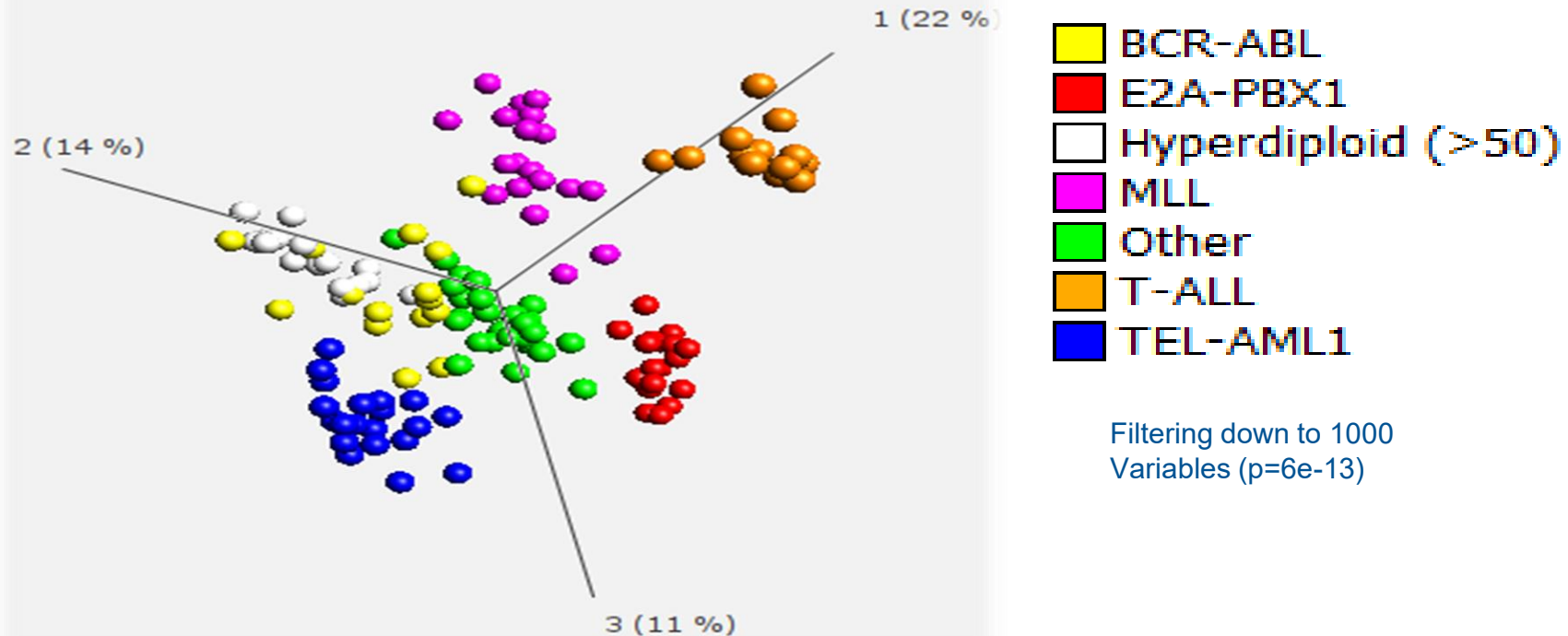
Dangers with exploratory analyses: ANOVA on Random data



Filtering using ANOVA on a 22282*132 dataset with significance threshold $p=0.05$ on random data resulting in 1108 “*significant discoveries*” + PCA visualization.

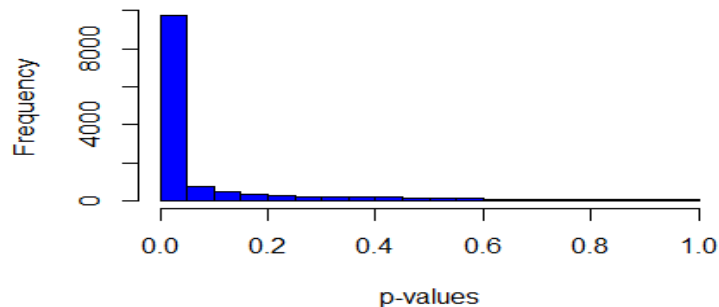
Note that $0.05 \times 22282 = 1114.1$

ANOVA on ALL dataset

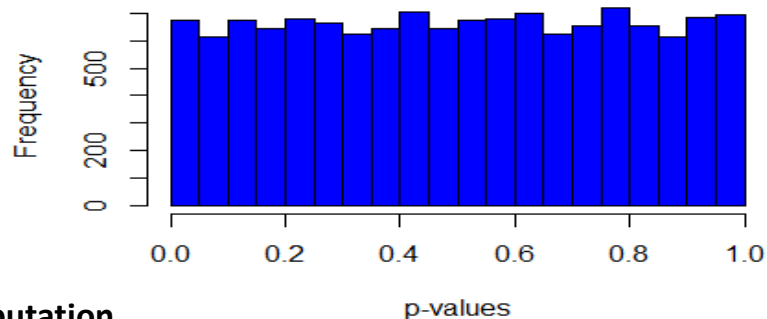


The p-value distributions

ANOVA on ALL dataset



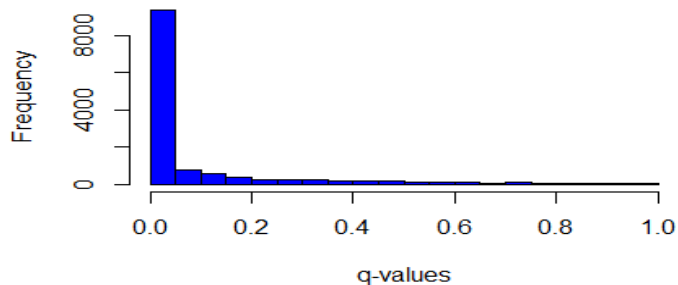
ANOVA on ALL randomized dataset



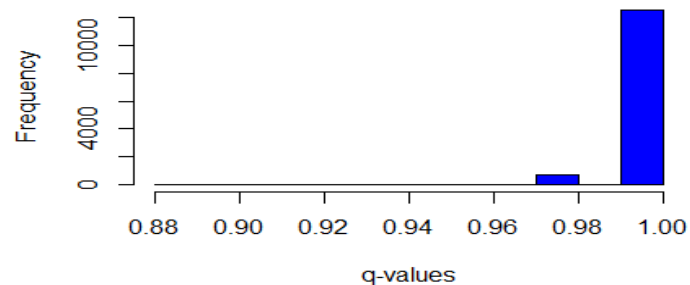
q-value computation



ANOVA on ALL dataset



ANOVA on ALL dataset randomized



I want to analyze the RNA expression data set on response to immune check point inhibition from the article <https://pubmed.ncbi.nlm.nih.gov/29443960/>. Please let me know what steps I should follow. I am interested in possibly alternative ways to look at the conclusions from the paper around TGF beta and response to cpi.

I'll ground this in the paper and the associated public data first, then outline an analysis plan with alternative tests of the TGF- β /CPI-response conclusion.

Thought for 12s >

The paper is the IMvigor210 atezolizumab metastatic urothelial cancer cohort. The core claim was that response was associated with CD8 T-effector biology and high neoantigen/TMB, while lack of response was associated with a fibroblast/TGF- β signature, especially in immune-excluded tumors. The EGA record says whole-transcriptome profiles were generated for 368 patients, and the RNA-seq dataset `EGAD00001003977` contains 34k FFPE tumor samples sequenced with Illumina TruSeq RNA Access / HiSeq 2500 before anti-PD-L1 treatment. [EGA Archive +1](#)

Here is a practical analysis plan.

1. Get the data in the most usable form

Start with processed data unless you specifically need raw FASTQs/BAMs.

The easiest routes are:

1. IMvigor210CoreBiologies R package

This package was created to reproduce analyses from the Mariathasan et al. manuscript and includes data/code for the paper. [GitHub](#)

2. Bioconductor [easierData](#)

This provides an exemplary bladder cancer dataset with processed RNA-seq data and ICB response metadata from Mariathasan et al. [Bioconductor +1](#)

3. Controlled access through EGA

Use this if you need the full original expression files and/or matched WES. The RNA dataset is `EGAD00001003977`; access is via the IMvigor210 data access committee. [EGA Archive](#)

For exploratory reanalysis, I would begin with `IMvigor210CoreBiologies` or `easierData`, then move to EGA only if required.

2. Reproduce the paper's baseline result first

Before doing alternative analyses, reproduce the headline figures/statistics:

- Define response as CR/PR versus SD/PD, matching the paper's approach where possible.
- Compute or retrieve:
 - CD8 T-effector score
 - TGF- β /fibroblast response score



Please provide R code for principal component analysis



Sure! Here is an example of how to perform principal component analysis (PCA) in R:

```
# Load data
data <- read.csv("data.csv")

# Perform PCA
pca <- prcomp(data, scale = TRUE)

# Print summary of results
summary(pca)

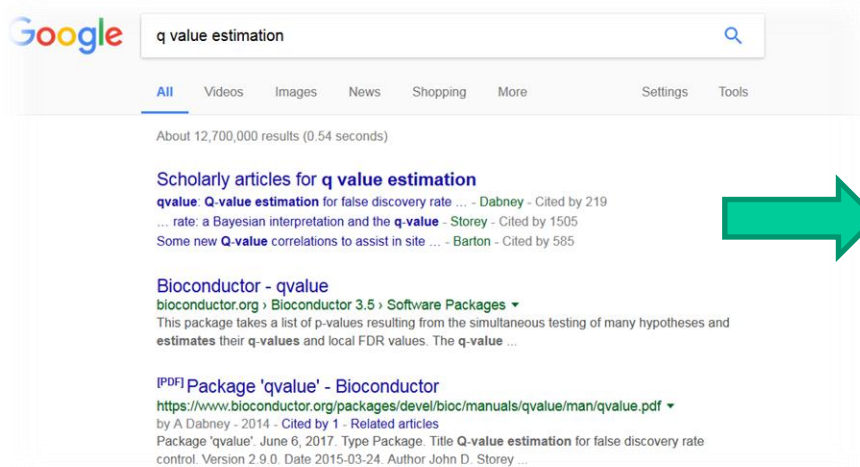
# Plot results
plot(pca, type = "l")
```



In this example, `data` is a dataframe containing the data we want to perform PCA on. We use the `prcomp()` function to perform the PCA, with the `scale = TRUE` argument indicating that we want to scale the data before performing the analysis. We then use `summary()` to print a summary of the results, and `plot()` to visualize the results.

Note that in practice, you may want to customize the PCA analysis to fit your specific needs. For example, you may want to specify the number of principal components you want to extract, or you may want to use a different scaling method. The `prcomp()` function has several optional arguments that allow you to customize the analysis, so be sure to consult the documentation for more information.

Google search for *q value estimation*



Google search results for "q value estimation". The search bar shows "q value estimation" and the results page displays "About 12,700,000 results (0.54 seconds)". The first result is "Scholarly articles for q value estimation" with a snippet about "qvalue: Q-value estimation for false discovery rate ... - Dabney - Cited by 219". A green arrow points from this result towards the Bioconductor page.

lead to R implementation in
Bioconductor →

[Home](#)[Install](#)[Help](#)

[Home](#) » [Bioconductor 3.5](#) » [Software Packages](#) » [qvalue](#)

qvalue

platforms **all** downloads **top 5%** posts **2 / 1 / 0.5 / 0** in Bioc **> 12 years**
build **ok** commits **0.33** test coverage **unknown**



Q-value estimation for false discovery rate control

Bioconductor version: Release (3.5)

This package takes a list of p-values resulting from the simultaneous testing of many hypotheses and estimates their q-values and local FDR values. The q-value of a test measures the proportion of false positives incurred (called the false discovery rate) when that particular test is called significant. The local FDR measures the posterior probability the null hypothesis is true given the test's p-value. Various plots are automatically generated, allowing one to make sensible significance cut-offs. Several mathematical results have recently been shown on the conservative accuracy of the estimated q-values from this software. The software can be applied to problems in genomics, brain imaging, astrophysics, and data mining.

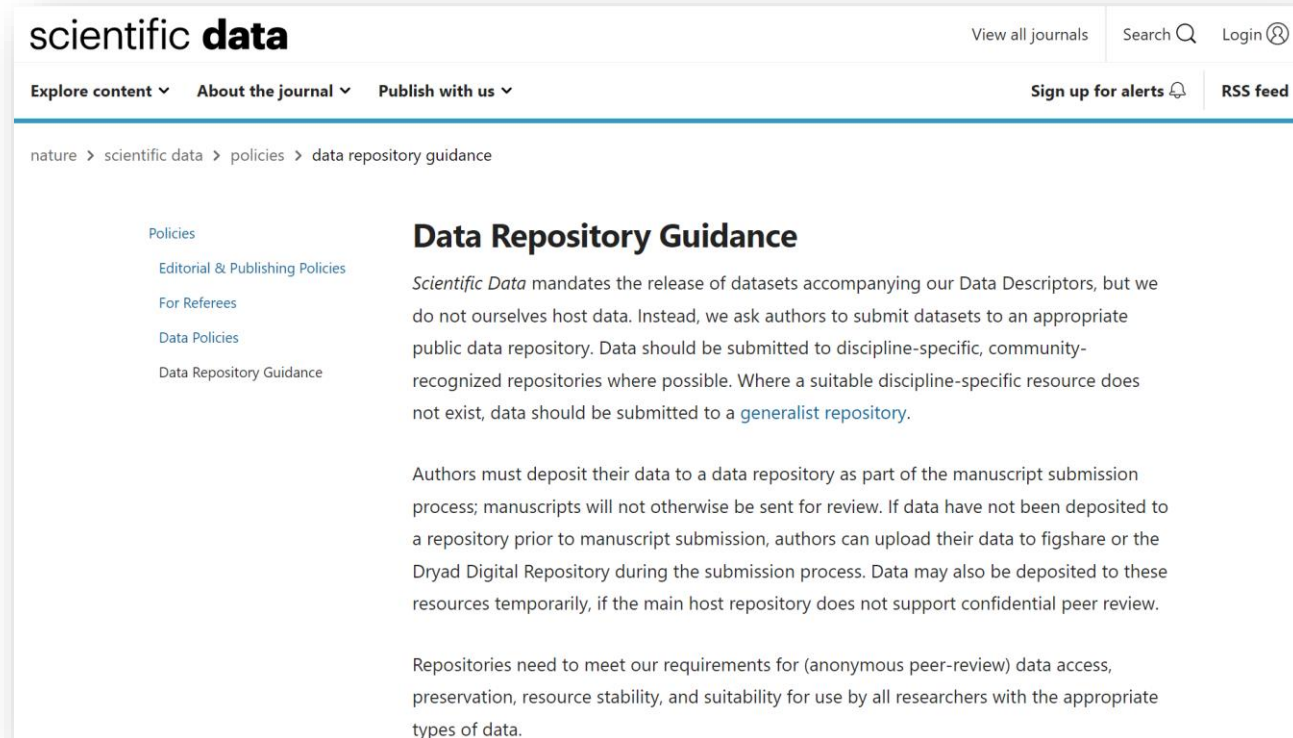
Author: John D. Storey with contributions from Andrew J. Bass, Alan Dabney and David Robinson

Maintainer: John D. Storey <jstorey at princeton.edu>, Andrew J. Bass <ajbass at princeton.edu>

Citation (from within R, enter `citation("qvalue")`):

Bass JDSwcfAJ, Dabney A and Robinson D (2015). *qvalue: Q-value estimation for false discovery rate control*. R package version 2.8.0, <http://github.com/jdstorey/qvalue>.

Where to find and deposit biomedical data?



The screenshot shows the 'scientific data' journal website. The header includes the journal name, navigation links like 'View all journals', 'Search', and 'Login', and a secondary navigation bar with 'Explore content', 'About the journal', and 'Publish with us'. Below this is a breadcrumb trail: 'nature > scientific data > policies > data repository guidance'. On the left, a sidebar lists 'Policies' with sub-links: 'Editorial & Publishing Policies', 'For Referees', 'Data Policies', and 'Data Repository Guidance'. The main content area is titled 'Data Repository Guidance' and contains three paragraphs of text regarding data deposition requirements and repository standards.

scientific **data** [View all journals](#) [Search](#) [Login](#)

[Explore content](#) [About the journal](#) [Publish with us](#) [Sign up for alerts](#) [RSS feed](#)

[nature](#) > [scientific data](#) > [policies](#) > [data repository guidance](#)

Policies

- [Editorial & Publishing Policies](#)
- [For Referees](#)
- [Data Policies](#)
- [Data Repository Guidance](#)

Data Repository Guidance

Scientific Data mandates the release of datasets accompanying our Data Descriptors, but we do not ourselves host data. Instead, we ask authors to submit datasets to an appropriate public data repository. Data should be submitted to discipline-specific, community-recognized repositories where possible. Where a suitable discipline-specific resource does not exist, data should be submitted to a [generalist repository](#).

Authors must deposit their data to a data repository as part of the manuscript submission process; manuscripts will not otherwise be sent for review. If data have not been deposited to a repository prior to manuscript submission, authors can upload their data to figshare or the Dryad Digital Repository during the submission process. Data may also be deposited to these resources temporarily, if the main host repository does not support confidential peer review.

Repositories need to meet our requirements for (anonymous peer-review) data access, preservation, resource stability, and suitability for use by all researchers with the appropriate types of data.

<https://www.nature.com/sdata/policies/repositories>

<https://www.ebi.ac.uk/>

EMBL-EBI home Services Research Training About us EMBL-EBI

EMBL's European Bioinformatics Institute

EMBL-EBI

Unleashing the potential of big data in biology

Find a gene, protein or chemical [Search](#)

Example searches: blast keratin bfl1 | About EBI Search

[Find data resources](#) [Submit data](#) [Explore our research](#) [Train with us](#)

Latest news



Genomic medicine: ushering in a new era in healthcare

08 Jun 2023



Janet Thornton retires: a pioneer in structural bioinformatics

07 Jun 2023



ProtVar: understanding missense variation in humans

02 Jun 2023

EMBL-EBI home Services Research Training About us EMBL-EBI

Services

Data resources and analysis tools to support life science research

EMBL's European Bioinformatics Institute (EMBL-EBI) maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.

Find a data resource or a tool [Search](#)

[Explore all our data resources and tools](#)

Our full range of data resources and data analysis tools are essential for supporting life science research.

Featured data resources

	AlphaFold DB Database for protein structure predictions for numerous species Web API CC-BY
	BioModels A repository of peer-reviewed, published, computational models. Web API CC0
	ChEMBL An open data resource of binding, functional and ADMET bioactivity data. Web API CC-BY

Featured tools

<https://www.ebi.ac.uk/services>



Expression Atlas

Gene expression across species and biological conditions

Query single cell expression

[To Single Cell Expression Atlas >](#)

- [Home](#)
- [Browse experiments](#)
- [Download](#)
- [Release notes](#)
- [FAQ](#)
- [Help](#)
- [Licence](#)
- [About](#)
- [Support](#)

Search across **66 species**, **4,451 studies**, **157,607 assays**

Ensembl 104, Ensembl Genomes 51, WormBase ParaSite 15, EFO 3.10.0

Search

Gene / Gene properties

Examples: [REG1B](#), [zinc finger](#), [O14777 \(UniProt\)](#), [GO:0010468 \(regulation of gene expression\)](#)

Species

Biological conditions

Examples: [lung](#), [leaf](#), [valproic acid](#), [cancer](#)

[Animals](#)
[Plants](#)
[Fungi](#)

<https://www.ebi.ac.uk/gxa/home>



[Homo sapiens](#)

1598 experiments

Baseline: 107



[Mus musculus](#)

1268 experiments

Baseline: 65



[Rattus norvegicus](#)

185 experiments

Baseline: 14



[Drosophila melanogaster](#)

149 experiments

Baseline: 5



[Sus scrofa](#)

45 experiments

Baseline: 16



[Gallus gallus](#)

39 experiments

Baseline: 4



[Login](#) [Register](#) [Need Help?](#)

[ABOUT](#) [DISCOVERY](#) [SUBMISSION](#) [ACCESS](#)

🔍 Search...

The European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of personally identifiable genetic, phenotypic, and clinical data generated for the purposes of biomedical research projects or in the context of research-focused healthcare systems.

I want to...



<https://ega-archive.org/>

https://www.ncbi.nlm.nih.gov/

 An official website of the United States government [Here's how you know](#) 



National Library of Medicine
National Center for Biotechnology Information

Log in

All Databases 

Search

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts into NCBI databases



Download

Transfer NCBI data to your computer



Learn

Find help documents, attend a class or watch a tutorial



Develop

Use NCBI APIs and code libraries to build applications



Analyze

Identify an NCBI tool for your data analysis task



Research

Explore NCBI research and collaborative projects



Popular Resources

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

[BLAST](#)

[Nucleotide](#)

[Genome](#)

[SNP](#)

[Gene](#)

[Protein](#)

[PubChem](#)

NCBI News & Blog

Now Available! Access Data from the Human Pangenome Research Consortium (HPRC) at NCBI

15 Jun 2023

[Have you ever wondered how your](#)

Making Discoveries in Canine & Human Oncology using the NIH Comparative Genomics Resource (CGR)

14 Jun 2023

[Do you use model organisms to study](#)

New! May 2023 Release of Stand-Alone



SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started

[How to Submit](#)[How to search and download](#)[How to use SRA in the cloud](#)[Submit to SRA](#)

Tools and Software

[Download SRA Toolkit](#)[SRA Toolkit Documentation](#)[SRA-BLAST](#)[SRA Run Browser](#)[SRA Run Selector](#)

Related Resources

[Submission Portal](#)[dbGaP Home](#)[BioProject](#)[BioSample](#)



COVID-19 Information

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)



Search for [Search](#) [Clear](#) [Show All](#) [Advanced Search](#)

Page size

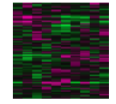
4348 DataSet records Page of 218 [>](#) [>>](#)

DataSet	Title	Organism(s)	Platform	Series	Samples
GDS6063	Influenza A effect on plasmacytoid dendritic cells	<i>Homo sapiens</i>	GPL10558	GSE68849	10
GDS6010	Influenza virus H5N1 infection of U251 astrocyte cell line: time course	<i>Homo sapiens</i>	GPL6480	GSE66597	18
GDS5879	Pulmonary CDC11c+ cells from young and middle-age animals	<i>Mus musculus</i>	GPL6885	GSE71868	8
GDS5826	Multiple myeloma cell lines with acquired resistance to chemotherapeutic agent carfilzomib	<i>Homo sapiens</i>	GPL570	GSE69078	12
GDS5825	Interleukin-1a deficiency effect on injured spinal cord	<i>Mus musculus</i>	GPL6246	GSE70302	12
GDS5881	Nebulin deficiency effect on the soleus	<i>Mus musculus</i>	GPL6246	GSE70213	12
GDS5880	Nebulin deficiency effect on the quadriceps	<i>Mus musculus</i>	GPL6246	GSE70213	12
GDS5913	SRPIN803 small molecule inhibitor of SRPK1 effect on retinal pigment epithelial cell line	<i>Homo sapiens</i>	GPL570	GSE62947	6
GDS5665	Pathogen-associated molecular-pattern curdian effect on interleukin-2 deficient GM-CSF myeloid dendritic cells	<i>Mus musculus</i>	GPL6246	GSE58120	12
GDS5662	Histone demethylase KDM3A-deficiency effect on estrogen-stimulated breast cancer cells in vitro	<i>Homo sapiens</i>	GPL10558	GSE68018	11

DataSet Record GDS6063: [Expression Profiles](#) [Data Analysis Tools](#) [Sample Subsets](#)

Title:	Influenza A effect on plasmacytoid dendritic cells		
Summary:	Analysis of primary plasmacytoid dendritic cells (pDC) exposed to influenza A for 8 hours ex vivo. pDCs are vital to antiviral defense, directing immune responses via secretion of interferon-alpha. Results provide insight into the regulation of the response of pDC to viral pathogens.		
Organism:	<i>Homo sapiens</i>		
Platform:	GPL10558: Illumina HumanHT-12 V4.0 expression beadchip		
Citation:	Bajwa G, DeBerardinis RJ, Shao B, Hall B et al. Cutting Edge: Critical Role of Glycolysis in Human Plasmacytoid Dendritic Cell Antiviral Responses. <i>J Immunol</i> 2016 Mar 1;196(5):2004-9. PMID: 26826244		
Reference Series:	GSE68849	Sample count:	10
Value type:	count	Series published:	2016/02/01

Cluster Analysis



Download

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINIML file
- Annotation SOFT file

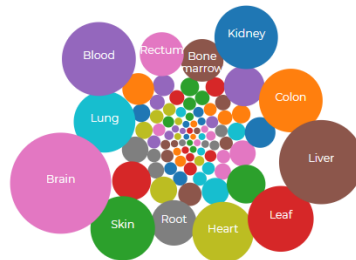
Organism, repository, gene, tissue, accession

Advanced Search

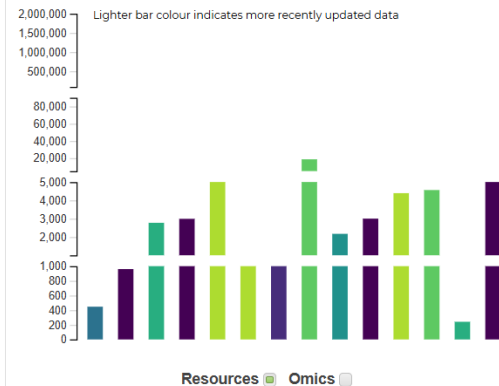
Examples: Cancer, Homo sapiens, Orbitrap, Q9HAU5, Phospho, Hela, PXD001416

more known regulation mechanisms dataset cellular related sequencing pathways potential including novel patients expression tumor mass molecular model disease derived analysis revealed further ms samples target through important

Description Sample Data



Tissues Organisms Diseases



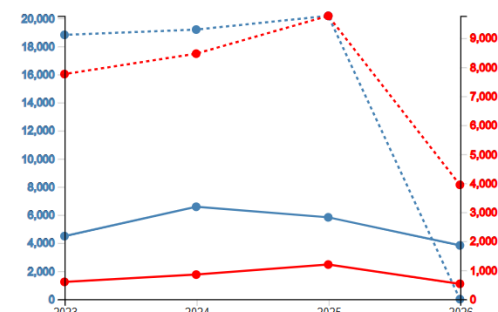
Latest Datasets

- P May 31 '26 Effects of different light conditions in Centella a...
- P May 31 '26 4a€²-O-(L-Alanylated) DMC as a EEB3 Inhibito...
- T May 30 '26 Implant renal injury-responsive cells to suppl...
- T May 30 '26 Multiomic profiling of vascular endothelial cell ...
- T May 30 '26 SoFIC Orchestrates Physical Barriers and Meta....
- G May 30 '26 Mouse MafA Cut&Run in Min6 cells
- T May 30 '26 CMPortal: a community-oriented, data-driven r...
- T May 30 '26 Sorbate induces lysine sorbylation through no...
- T May 30 '26 Virus-inclusive single-nucleus RNA sequencing...
- T May 30 '26 Single-nucleus RNA sequencing elucidates the...

Most Accessed Datasets

- SB 15787 Kholodenko1999 - EGFR signaling
- SB 12772 Ndairou2020 - early-stage transmission dynamics ...
- M 9440 A metabolomic study of urinary changes in type 2 ...

New Datasets Per Year



<https://www.import.org/home>

 ImmPort
  Upload
  Shared
  Analysis
  Resources
  Data Management and Sharing

 Documentation
  About



ImmPort is funded by the NIH, NIAID and DAIT in support of the NIH mission to share data with the public. Data shared through ImmPort has been provided by NIH-funded programs, other research organizations and individual scientists ensuring these discoveries will be the foundation of future research.

Data uploading or sharing questions? Please contact ImmPort_Helpdesk@immport.org.





Upload Data

Online Study Registration **(New)**

Upload Templates

Search Private Data



Shared Data

Data Model

Search/Download

Gene Lists

News & Events

06/18/2024 -  ImmPort Big Data in Immunology Workshop at FOCiS 2024 presented by Sanchita Bhattacharya - VirtualSan Francisco, CA 

06/12/2024 -  ImmPort Data Release 51.3 is out! 1 new study. For details please see the Data Release notes. 

05/23/2024 -  ImmPort Data Release 51.2 is out! 36 new studies, 14 updated studies. For details please see the Data Release notes. 

Provide feedback

[ImPort](#)
[Shared Data](#)
[Analysis](#)
[Resources](#)
[Data Management and Sharing Plan](#)
[Search](#)
[Data >](#)
[About >](#)

[Shared Data](#)
[Data Catalogue](#)
[Data Model](#)
[Help >](#)

Search studies will now be findable via an intuitive new search interface with ImmPort.
 View new documentation about accessing longitudinal/ COVID-19 data from the latest IMPACC study.
 Have data to share? Try the online wizard for quick and easy study registration within ImmPort.
 Shared Data quick links:
 [COVID-19 studies](#)
[Influenza studies](#)
[Respiratory-like illnesses studies](#)
[Viral infectious diseases studies](#)



IMMPORT
Shared Data
your data for research and sharing

- Browse Shared Data using Faced Search
- Download Data (Login Required)
- API Documentation
- Ask ImmPort Questions

Program specific resource links:
 [AMP RABLE](#)
[Bill & Melinda Gates Foundation](#)
[HPC](#)
[March of Dimes](#)
[Sealcoat](#)

Enter text (or) Click 'Search' to explore ImmPort shared data e.g. influenza, COVID-19, SDY1...
 [Search](#)

[Data Summary: Release 48, May 2023](#)

[Click on the counts with a icon to visualize the count breakdown](#)

Studies

121	91877	148
-----	-------	-----

Experiments

2686	6528595	1279770
------	---------	---------

Subjects

Diseases

[Click on a bubble to zoom in](#)
[Research Focus by Assay Type](#)

[Bubble Summary: Research Focus by Assay Type](#)

<https://www.immport.org/shared/home>

https://www.broadinstitute.org/



[Careers](#) [Giving](#) [Intranet](#) [Q](#)

[About us](#) ▾ [Research](#) ▾ [Centers](#) ▾ [Education and outreach](#) ▾ [News](#) ▾

[HOME](#) | [TECHNOLOGY AREAS](#)

Data Sciences Platform

We build technology to help researchers connect to the patients, datasets and tools they need to do life-changing biomedical research.

The life sciences are in the midst of a data revolution. Cheap and accurate genome sequencing is a reality, advanced imaging is routine, and clinical data is increasingly stored in electronic formats. These innovations — and the massive data sets they produce — have brought us to the threshold of a new era in medicine, one where the data sciences hold the potential to propel our understanding and treatment of human disease.

The Broad Data Sciences Platform (DSP) is a methods development and software engineering group dedicated to maximizing the impact of the data sciences on

AREAS OF FOCUS



[Workbench](#)

[Analytical Tools](#)

[User Interfaces](#)

[Production Data Processing](#)

[SPEAKER SERIES](#)



https://www.broadinstitute.org/data-sciences-platform

<https://terra.bio/> - A joint effort between the Broad Institute & Microsoft & Verily



Welcome to Terra Community Workbench

Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate.

Find how-to's, documentation, video tutorials, and discussion forums [↗](#)

Learn more about the Terra platform and our co-branded sites [↗](#)

View Workspaces

Workspaces connect your data to popular analysis tools powered by the cloud. Use Workspaces to share data, code, and results easily and securely.

[→](#)

View Examples

Browse our gallery of showcase Workspaces to see how science gets done.

[→](#)

Browse Data

Access data from a rich ecosystem in DUOS.

[→](#)

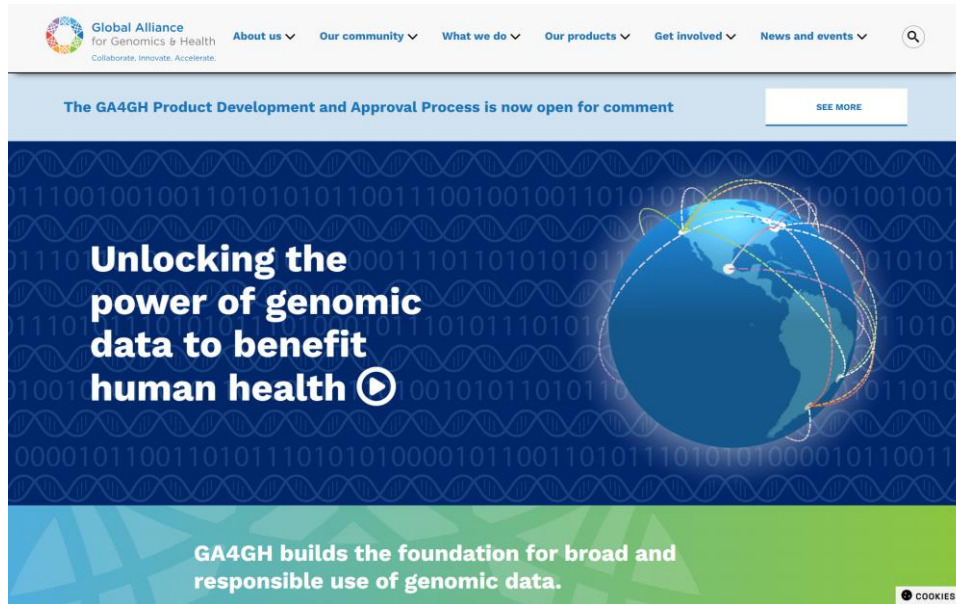
Terra Roadmap

Stay connected: see what's new, what's next, and try out the latest features on Terra.

[→](#)

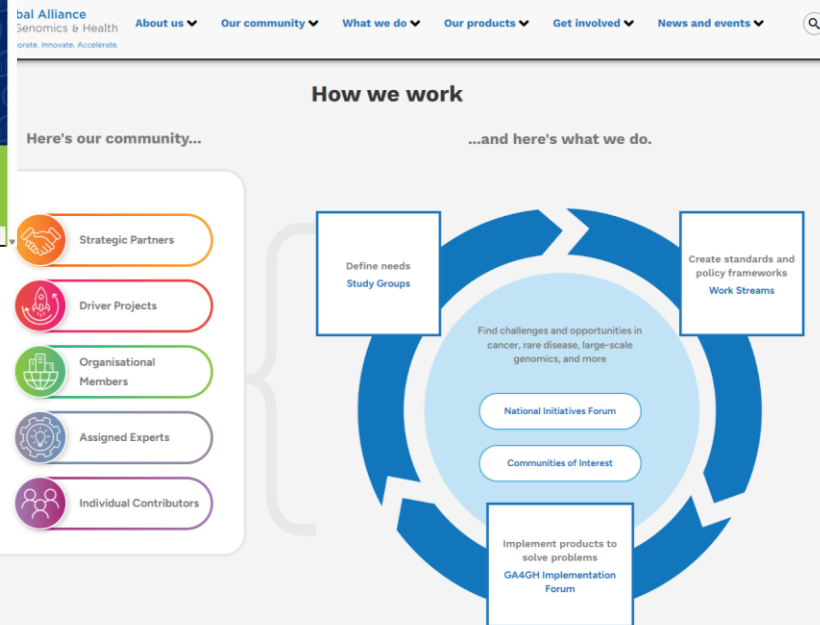


This project has been funded in whole or in part with Federal funds from the National Cancer

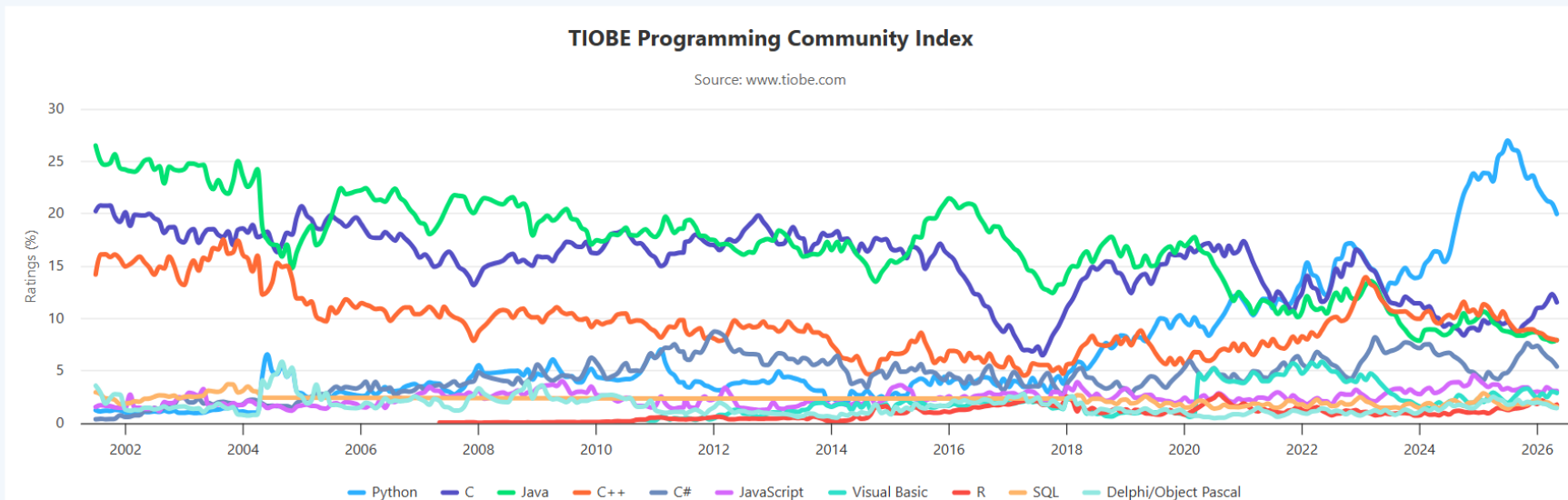


<https://www.ga4gh.org/>

Formed in 2013, the Global Alliance for Genomics and Health (GA4GH) unites an international community dedicated to advancing human health through genomic data. We build technical standards and policy frameworks and tools that will expand responsible, voluntary, and secure use of genomic and other related health data.



Interacting with Data - Software tools



<https://www.tiobe.com/tiobe-index/>

What tools do I use?

→ **Python** and **R** are still dominating in bioinformatics



[Home]

Download
CRAN

R Project
About

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

<https://www.r-project.org/>

The Posit website banner features a background image of a river flowing through a lush green landscape. The Posit logo is in the top left. Navigation links include About, Partners, Download RStudio, and Download Positron. A search bar and buttons for GET STARTED and DEMO GALLERY are in the top right. Below the navigation, there are buttons for BUSINESS LEADER and PRACTITIONER. The main text reads 'Turn data science into business impact.' followed by a sub-headline 'The enterprise platform for the AI era, giving data teams speed to build and IT the control to scale it safely.' On the right, there is a video player showing two men looking at a computer screen with a data visualization overlay.

posit

About Partners Download RStudio Download Positron

Solutions Products Open Source Customers Resources GET STARTED DEMO GALLERY

BUSINESS LEADER PRACTITIONER

TRUSTED BY 52 OF THE FORTUNE 100

Turn data science into business impact.

The enterprise platform for the AI era, giving data teams speed to build and IT the control to scale it safely.

About *Bioconductor*

The mission of the *Bioconductor* project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays. We are dedicated to building a diverse, collaborative, and welcoming community of developers and data scientists.

Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. *Bioconductor* is also available as [Docker](#) images.

<https://www.bioconductor.org/>

Install »

- Discover [2083 software packages](#) available in *Bioconductor* release 3.14.

Get started with *Bioconductor*

- [Install *Bioconductor*](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn »

Master *Bioconductor* tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use »

Create bioinformatic solutions with *Bioconductor*

- [Software](#), [Annotation](#), and

Develop »

Contribute to *Bioconductor*

- [Developer resources](#)
- [Use Bioc 'devel'](#)

Interactive data analysis using e.g. RShiny web-applications

<https://gdc.cancer.gov/>

NCI Cancer Research Data Commons

NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons

CCG Web Site | Contact Us | [Launch Data Portal](#) | GDC Apps

Search this website

About the GDC | About the Data | **Analyze Data** | Access Data | Submit Data | For Developers | Support | News

The Next Generation Cancer Knowledge Base

Cases by Major Primary Site



The NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified repository and cancer knowledge base that enables data sharing across cancer genomic studies in support of precision medicine.

The GDC supports several cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET).

[More about the GDC](#)

Analyze Data

The **GDC Data Analysis, Visualization, and Exploration (DAVE) Tools** allow users to interact intuitively with the GDC data and promote the development of a true cancer genomics knowledge base.

[More about Analyzing Data](#)

Access Data

The **GDC Data Portal** provides a platform for efficiently querying and downloading high quality and complete data. The GDC also provides a **GDC Data Transfer Tool** and a **GDC API** for programmatic access.

NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons

CCG Web Site | Contact Us | [Launch Data Portal](#) | GDC Apps

Search this website

About the GDC | About the Data | **Analyze Data** | Access Data | Submit Data | For Developers | Support | News

Analyze Data

The GDC provides user-friendly and interactive Data Analysis, Visualization, and Exploration (DAVE) Tools supporting gene and variant level analysis that allows researchers to:

- ✓ Visualize most frequently mutated genes and view most frequent somatic mutations for a project
- ✓ Plot all cases for a project in an OncoGrid and visualize the top 50 mutated genes affected by high impact mutations
- ✓ Perform a survival analysis for cases with a mutated form of a certain gene and cases without the mutation
- ✓ Visualize mutations and their frequency across cases mapped to a graphical visualization of protein-coding regions using an interactive Protein Viewer
- ✓ View the cancer distribution as evidenced by the number of cases affected by the mutation across all projects
- ✓ Build cohorts and perform gene and variant level analysis on the cohort
- ✓ Compare custom gene or case sets by visualizing set similarities and differences

[Get Started by Exploring GDC Data Analysis Processes and Tools »](#)



<https://gdc.cancer.gov/analyze-data>

<https://cri-iatlas.org/>

CRI iAtlas Portal

Explore

Documentation

Notebooks

About

Resources

Help

iAtlas Explorer Home

ICI Cohort Selection

ICI Analysis Modules

Datasets Overview

Clinical Outcomes

Hazard Ratio

Immune Features

Immunomodulators

Machine Learning

CG Cohort Selection

CG Analysis Modules

Cell-Interaction Diagram

Clinical Outcomes

CNV Associations

Driver Associations

Extracellular Networks

Germline Analysis

Immune Feature Trends

Immunomodulators

IO Targets

TIL Maps

Tumor Microenvironment

iAtlas tools

Immune Subtype Classifier

Data Description

iAtlas Explorer - Home

What's Inside

Explore immune response in tumor tissue samples from **Immune Checkpoint Inhibitor (ICI)** studies or **Cancer Genomics (CG)** studies, using CRI iAtlas Analysis Modules for interactive visualization.

Begin by selecting either a cohort from available studies on molecular response to ICI (**ICI Cohort Selection**) or a CG cohort (**CG Cohort Selection** for **TCGA** or **PCAWG**). You can also select **Data Description** on the left navigation bar to learn which immune readouts are available.

Immune Checkpoint Inhibitors (ICI) datasets: 15

Immune Readouts: 270

Cancer Genomics (CG) datasets: 2

Samples: 12886

Get Started

1. Build your Cohort

Use our cohort selector to explore the available data and narrow down your research targets.

Open ICI Cohort Selection

Open CG Cohort Selection

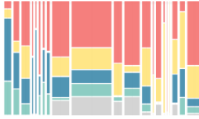
2. Visualize your data

Use our analysis modules to explore the selected cohorts. You can access the analysis modules from the sections below and from the left menu. Any changes in the selected cohort in step 1 will be automatically propagated to the corresponding modules.

Immune Checkpoint Inhibition Analysis Modules

Each module in this section provides visualizations of datasets with data from datasets from studies of treatments with Immune Checkpoint Inhibitors (ICI).

Datasets Overview



Explore categories and groups of the available datasets.

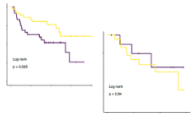
Open Module

Hazard Ratio



Create Cox Proportional Hazard Regression Models and visualize Hazard Ratio in a

Clinical Outcomes



Plot survival curves based on immune characteristics and identify variables associated with outcome.

Open Module

Immune Features



See how immune readouts vary across groups and ICI datasets.

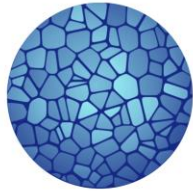
<https://isb-cgc.shinyapps.io/iatlas/>

Focusing on Cells – The basic units of Life



HUMAN
CELL
ATLAS

About HCA • Data • Research • Ethics and Equity • News/Events • Publications Resources •



HUMAN
CELL
ATLAS

<https://www.humancellatlas.org/>

HUMAN CELL ATLAS DATA PORTAL

Explore Guides Metadata Pipelines Analysis Tools Contribute APIs Updates

Mapping the Human Body
at the Cellular Level

Community generated, multi-omic, open data

44.5M CELLS 6.9k DONORS 364 PROJECTS 567 LABS

FIND PROJECTS Filter projects by attribute e.g. project title GO

This website uses cookies for security and analytics purposes. By using this site, you agree to these uses. Learn more here.

Get it

<https://data.humancellatlas.org/>

On June 15th 2023 Aviv Regev received the L’Oreal-Unesco for women in science award *“for her pioneering work applying mathematics and computer science to revolutionize cell biology.”*



MEET THE 2023 LAUREATES

L'ORÉAL-UNESCO FOR WOMEN IN SCIENCE
INTERNATIONAL AWARDS





APPLICATION

[Collections](#) [Datasets](#) [Gene Expression](#) [Cell Guide](#) [Differential Expression](#) **NEW**

CENSUS

[API](#) [Models](#)[Explore Your Data](#)[Help & Documentation](#)**CZ CELL × GENE**
DISCOVER

Discover the mechanisms of human health

Download and visually explore data to understand the functionality of human tissues at the cellular level with Chan Zuckerberg CELL by GENE Discover (CZ CELLxGENE Discover).

UNIQUE CELLS

156.1M

DATASETS

2124

CELL TYPES

1117[Differential Expression](#)[Explorer](#)[Census](#)[Cell Guide](#)[Collections & Datasets](#)[Gene Expression](#)[Annotate](#)

Omics Overview

Summary of All Tissues (Final Cell/Sample Counts)

<https://www.singlecellatlas.org/>



scRNA-Seq

Cell count: 3,881,472



scATAC

Cell count: 773,190



scImmune

Unique clonotypes: 167,379



Integrative scImmune &
scRNA-Seq

Cell count: 209,708



Spatial Transcriptomics

Tissues: 50 (fetal+adult)



RNA-Seq+Genotype

RNA-Seq samples: 16,704
Genotype samples: 837



CyTOF

Cell count: 2,278,550



Flow Cytometry

Cell count: 192,925,633



[Packages ▾](#)[Learn](#)[People](#)[Blog](#)[Events](#)[About ▾](#)[Join](#)[September 2024](#)[The very first scverse conference! →](#)

scverse

A library of
single cell
analysis
algorithms and
computational
models

scverse

Foundational tools for single-cell omics data
analysis

[GitHub](#)[Discourse](#)[Zulip](#)[Twitter](#)[YouTube](#)

```
import scanpy as sc
```



```
sc.pl.pca(adata)
```



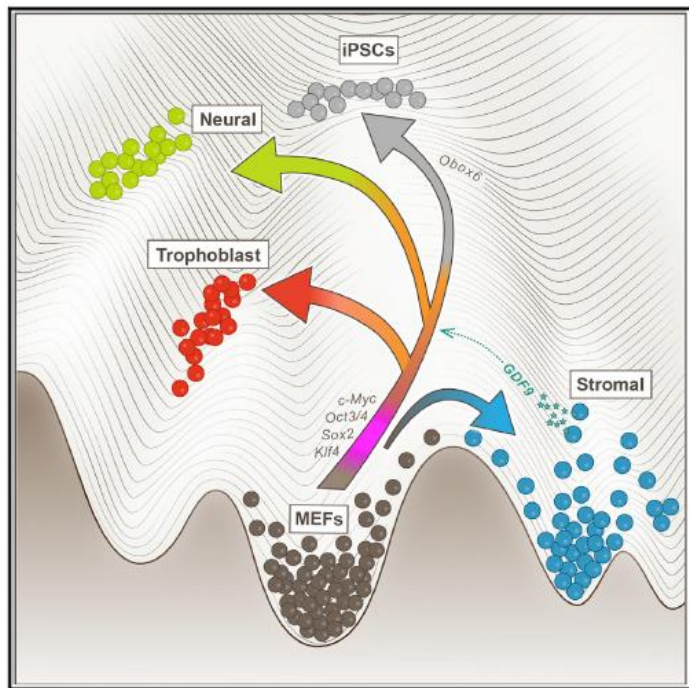
<https://scverse.org/>

MISSION

scverse is a consortium of foundational tools (mostly in Python) for omics data in life sciences. It has been founded to ensure the long-term maintenance of these core tools.

Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming

Graphical Abstract



Authors

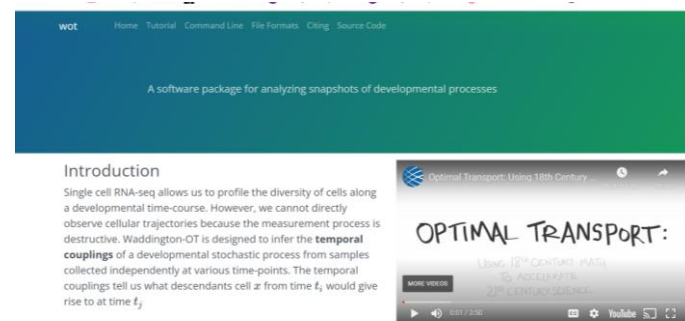
Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, ..., Rudolf Jaenisch, Aviv Regev, Eric S. Lander

Correspondence

jianshu@broadinstitute.org (J.S.), aregev@broadinstitute.org (A.R.), lander@broadinstitute.org (E.S.L.)

In Brief

Application of a new analytical approach to examine developmental trajectories of single cells offers insight into how paracrine interactions shape reprogramming.



<https://broadinstitute.github.io/wot/>