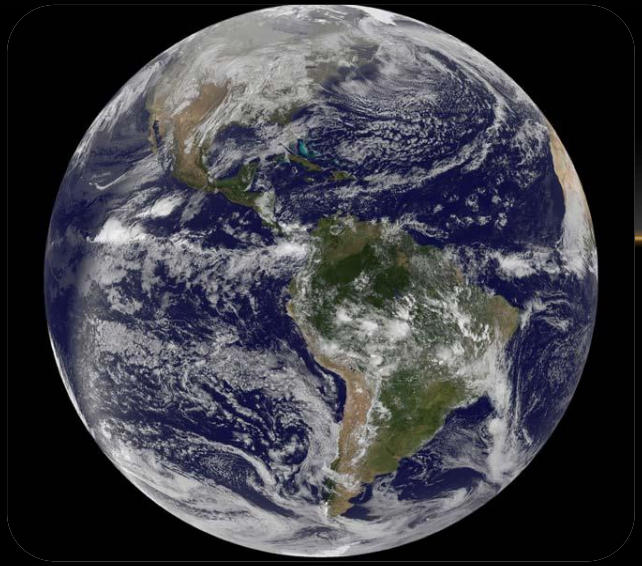
Systems Immunology & Immune Oncology
A Data-Centric View



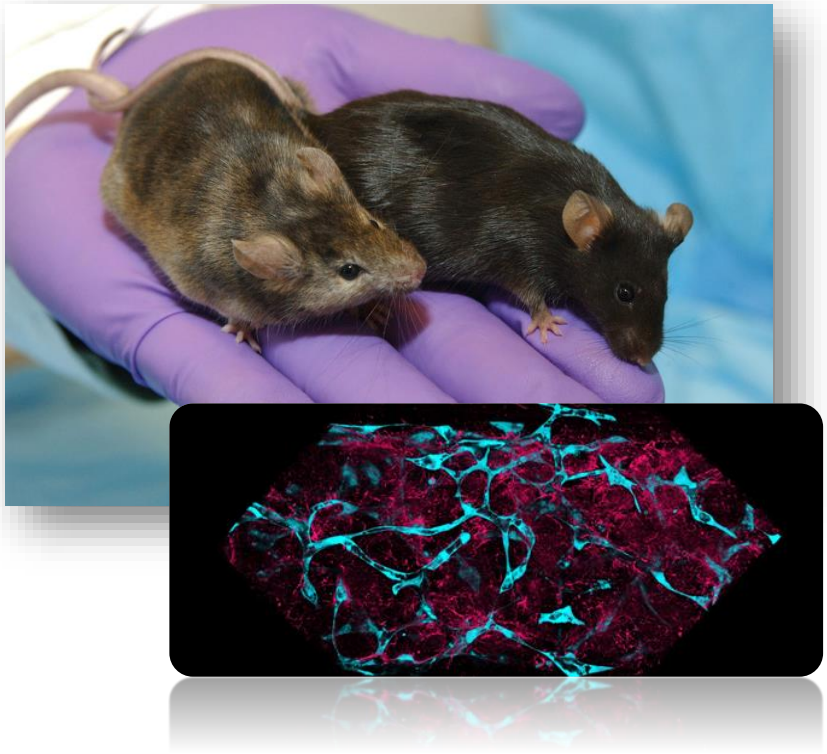
Magnus Fontes

https://pics-about-space.com/pics-new-world-nasa?p=2

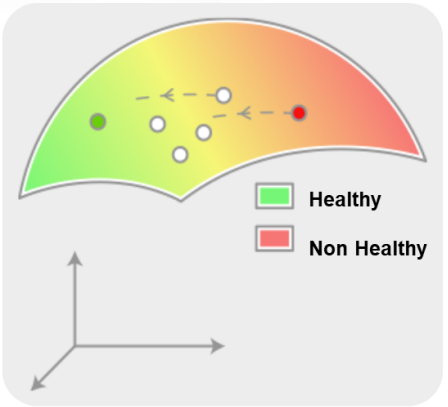
Disclosures & Affiliations

- 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.qlucore.com
- Some images and diagrams have been created partly using Gemini or ChatGPT.

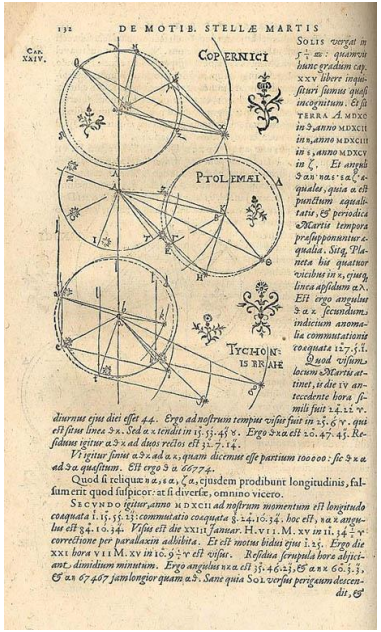
Multidisciplinary *models* to try to *understand* biology, But, what is a model?



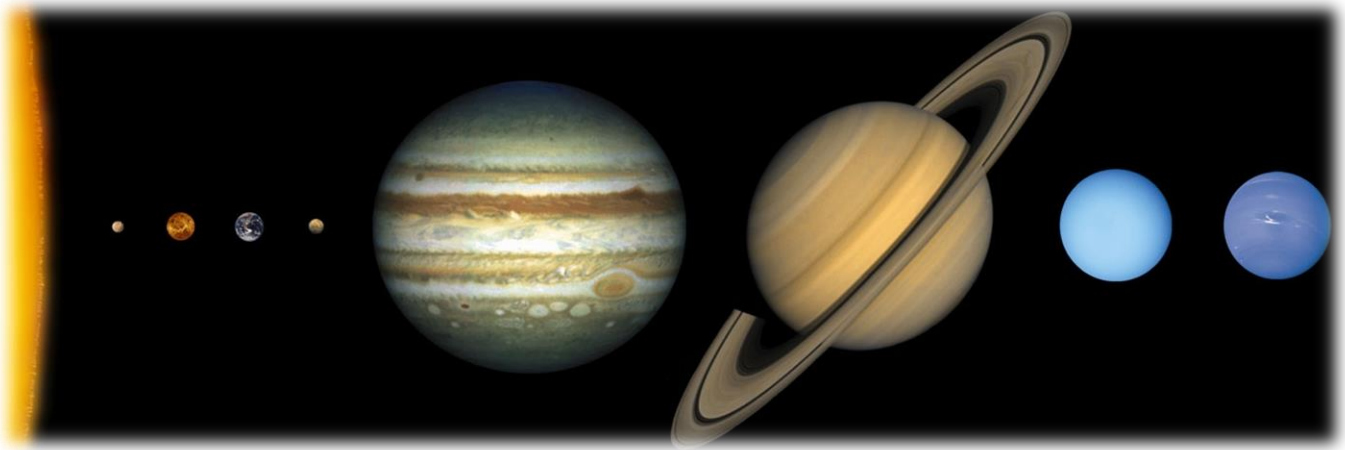
$$\frac{dx}{dt} = \alpha x - \beta xy,$$
$$\frac{dy}{dt} = \delta xy - \gamma y,$$



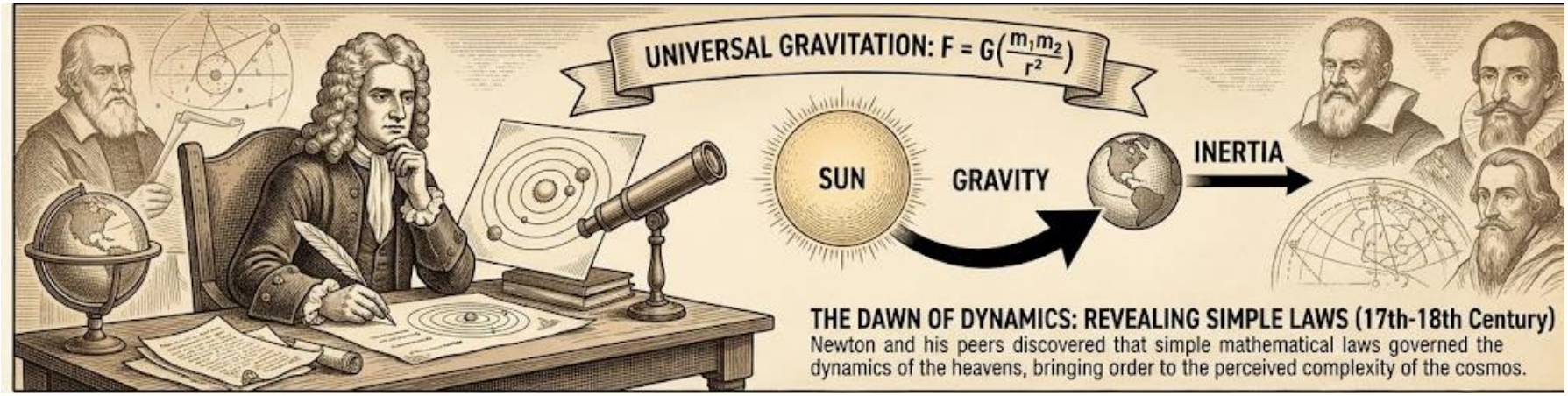
Modeling a famous System



Astronomia Nova,
Johannes Kepler, 1609



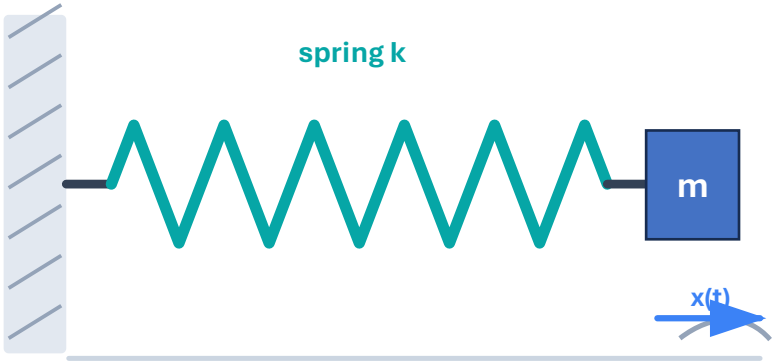
https://solarsystem.nasa.gov/multimedia/display.cfm?IM_ID=178



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

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

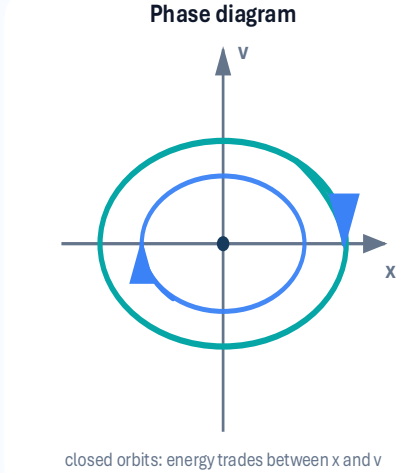
state-space dynamics

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

undamped intuition

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

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

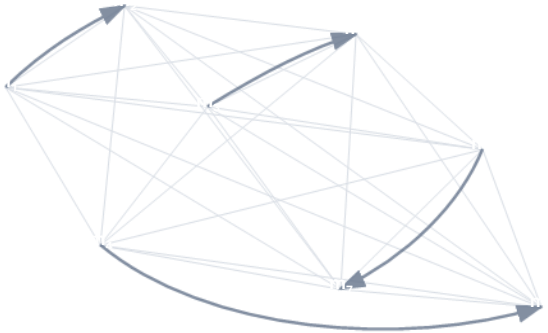


State space makes the prediction visual: one point contains the system’s future trajectory.

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

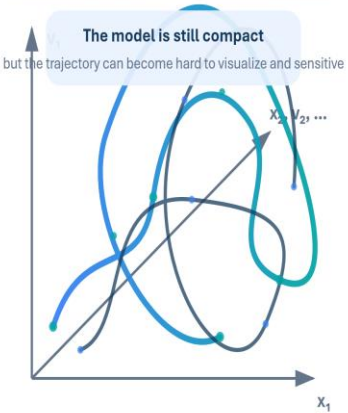
$$m_i \dot{\mathbf{v}}_i = \sum_{j \neq i} G\, m_i m_j \frac{\mathbf{r}_j - \mathbf{r}_i}{\|\mathbf{r}_j - \mathbf{r}_i\|^3}$$

state vector

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

complexity grows fast

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



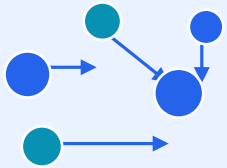
The “point” in state space now lives in many dimensions; projections look tangled and sensitive.

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

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

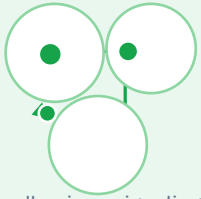
Known molecule–molecule interactions can be written compactly, but they propagate across scales into nonlinear, context-dependent tissue behavior.

Molecule–molecule



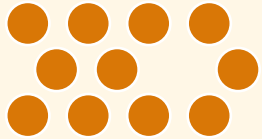
pair potentials
binding affinities
reaction rates

Cell–cell



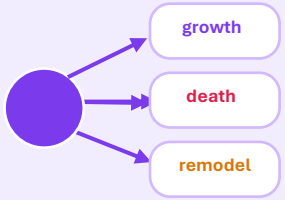
adhesion · signaling
mechanical contact

Tissue microenvironment



diffusion fields
stress gradients
spatial constraints

Exposed tissue response



same laws, different context
→ different phenotype

Compact equations at each scale

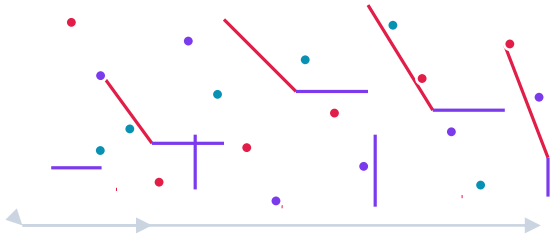
Molecules: ODEs to PDEs

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

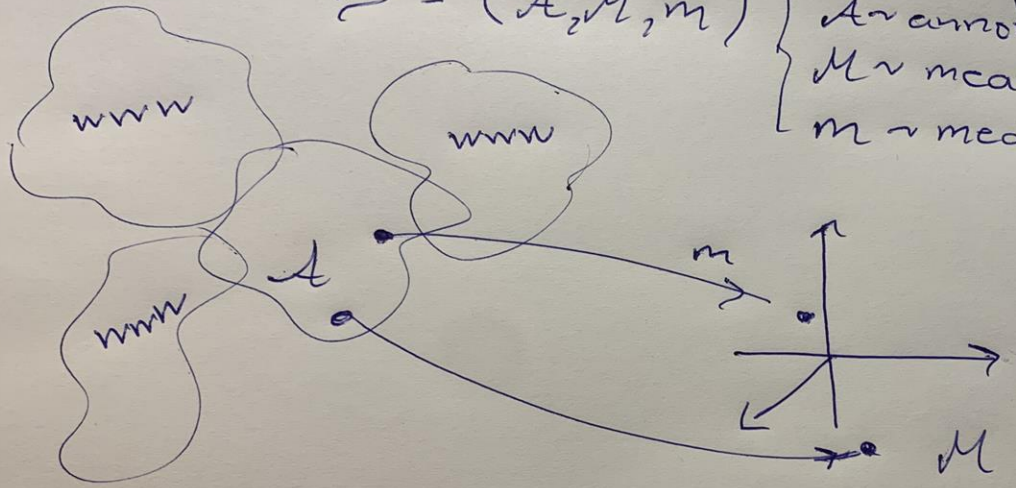


x = {molecular concentrations, receptor states, cell positions, phenotypes, tissue fields, environment…}

dimension ≫ 2, nonlinear, multimodal, hierarchical, stochastic

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

$\mathcal{D} = (\mathcal{A}, \mathcal{M}, m)$ { $\mathcal{A}$ ~ annotation space; $\mathcal{M}$ ~ measurement space; m ~ measurement map

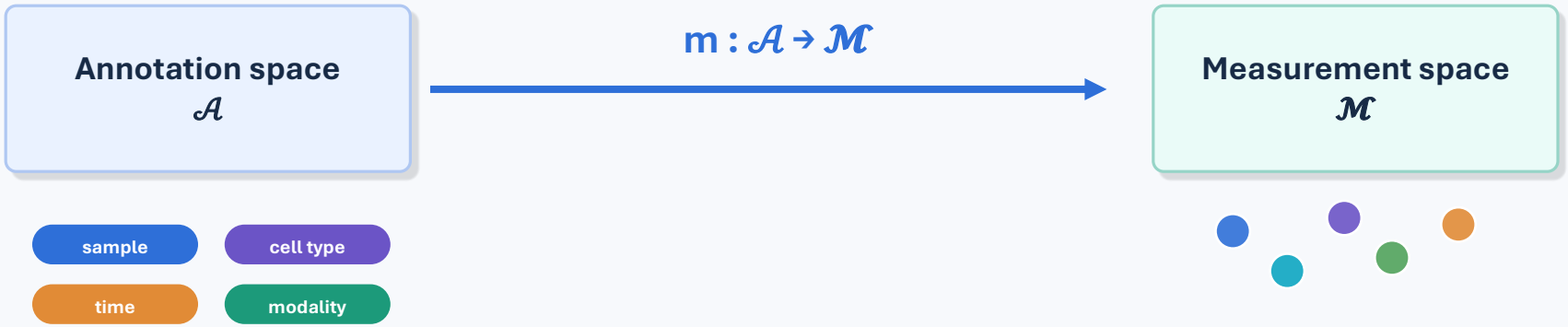


$\mathcal{A}$ is the realm of AI agents

$\mathcal{M}$ 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) -> 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

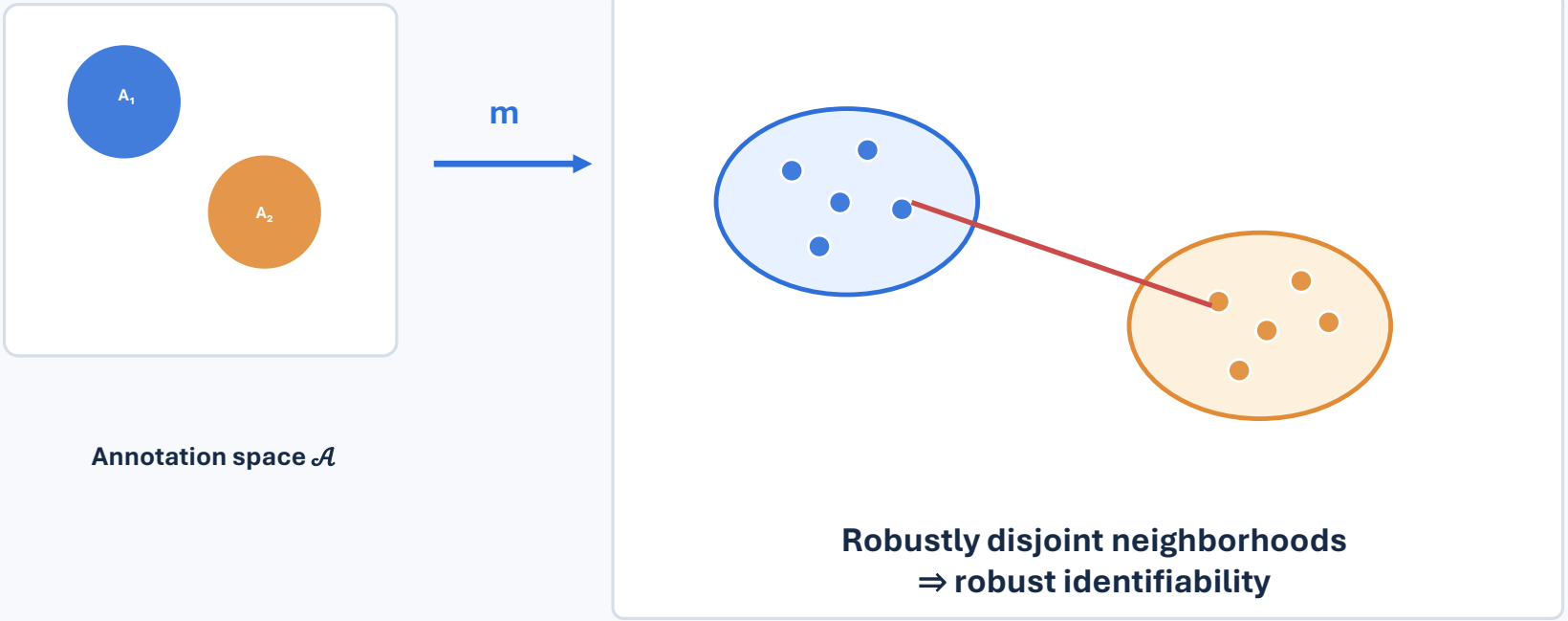
$\mathcal{A} = S \times T \times V \times \text{Space}$

fixed sample s and variable v
⇢ spatiotemporal field
biomarkers may be
shape, flow, or dynamics

Annotation structure induces vectors, matrices, trajectories, and fields

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.

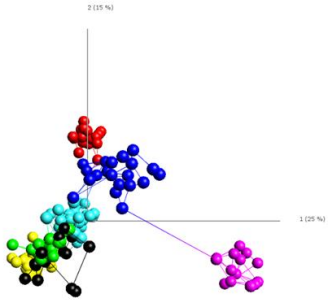


Geometry choices create different biomarker notions

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

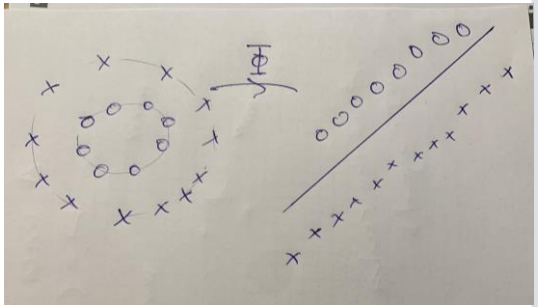
Graph geometry

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.

1. **Annotate** — Define biological domains A_i inside annotation space $\mathcal{A}$. This is the realm where LLMs excel
2. **Measure** — Map domains into measurement space: $X_i = m(A_i)$. Here there is room for traditional mathematical and statistical modeling.
3. **Choose geometry** — Metric, graph, Hilbert, kernel, path, or function space.
4. **Separate or mix** — Biomarkers separate X_i , based on geometry choice define robustness
5. **Assess robustness** — Use geometric ε-neighborhoods, statistical margins, sparsity to get robust identifiability

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

Modeling and controlling immunodynamics

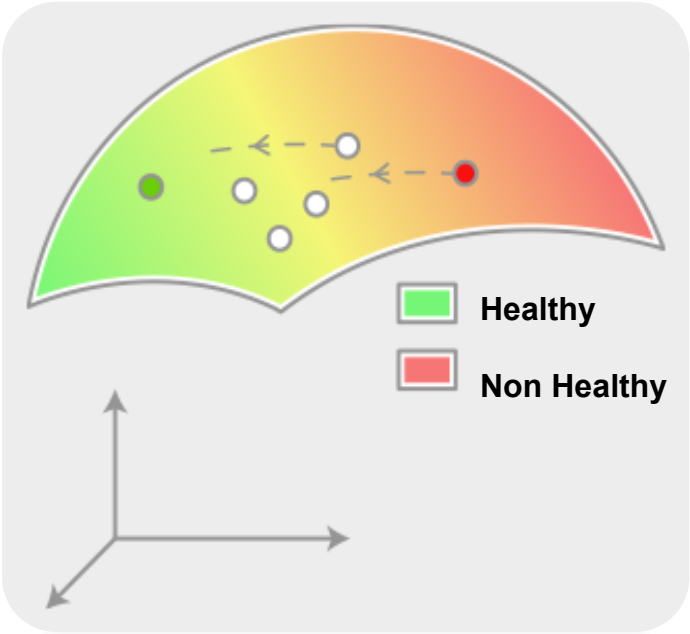
The Holy grail of modern biomedical research is to *predict and control* the human immune system

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**



Transformation of Biomedical Science, Research and Development

Challenges and Opportunities

“*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:

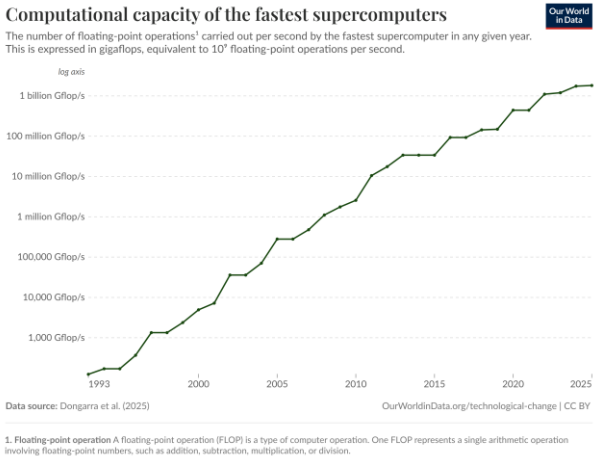
- Discovery of the DNA Double Helix structure (1953)
- Recombinant DNA and Protein Technologies (1970s onwards)
- Monoclonal Antibody Technology (1970s onwards)
- The Human Genome Project (completed 2003)
- CRISPR Gene Editing Technology (2012 onwards)
- Emergence of Gene and Cell Therapies

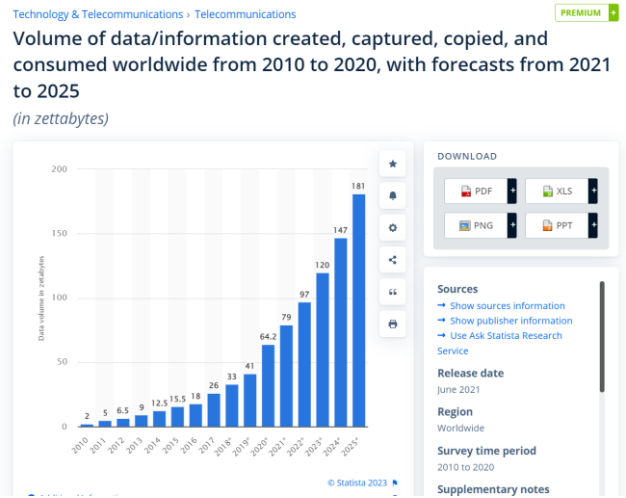
+ Proliferation of Advanced Measuring Technologies:

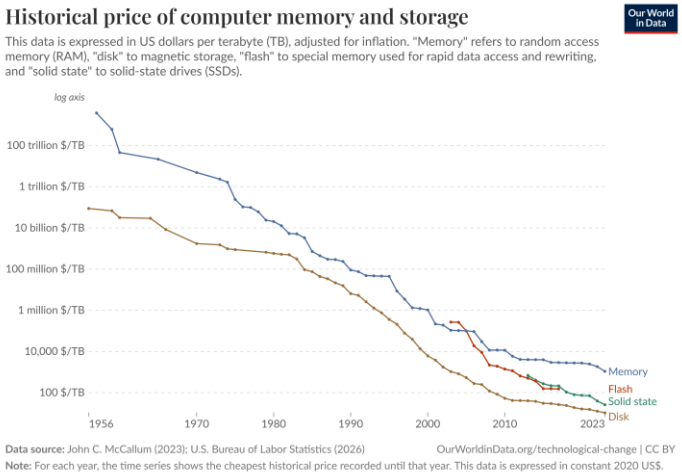
- New sequencing technologies (e.g., Next-Generation Sequencing)
- High-throughput measuring techniques (e.g., proteomics, metabolomics)

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

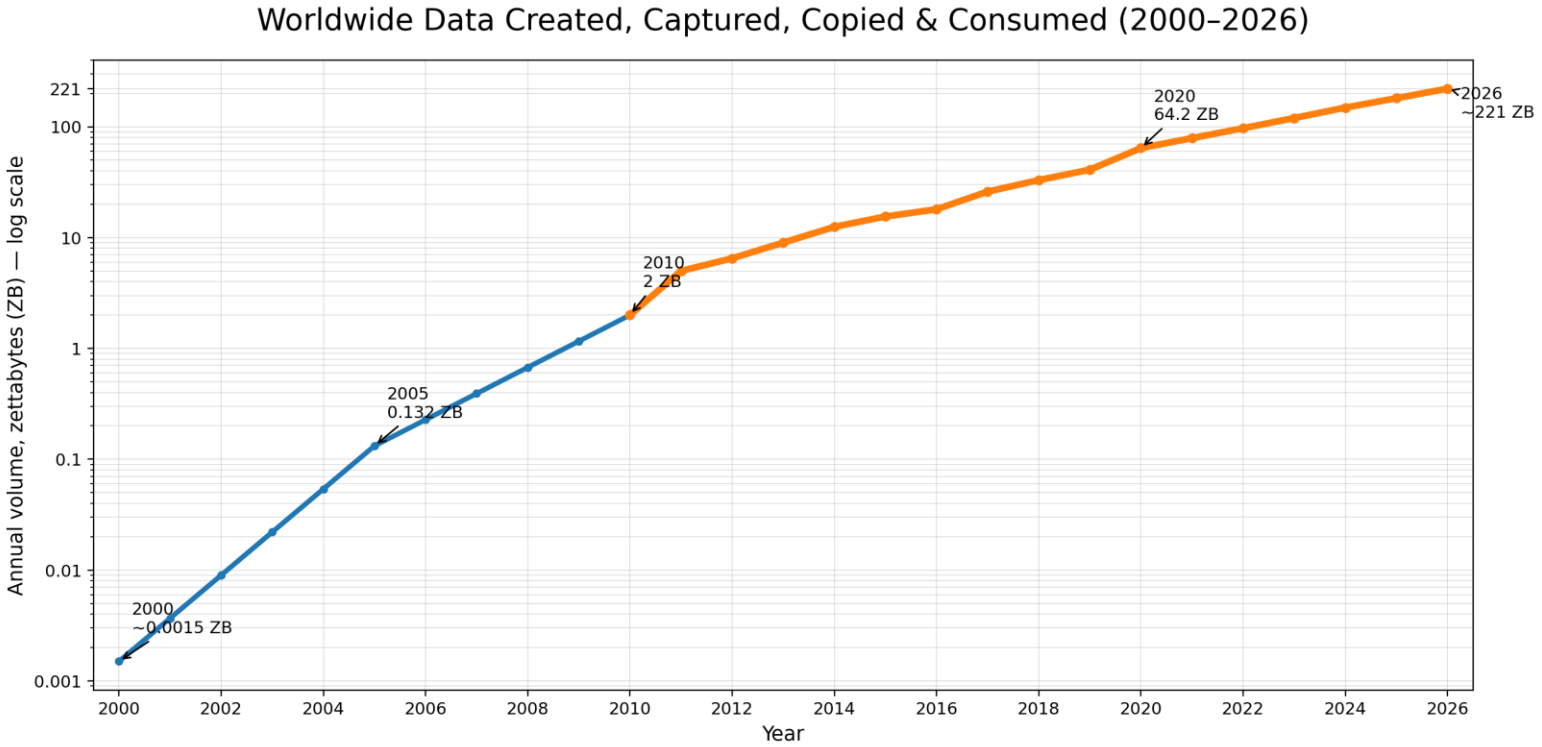
Moore's law







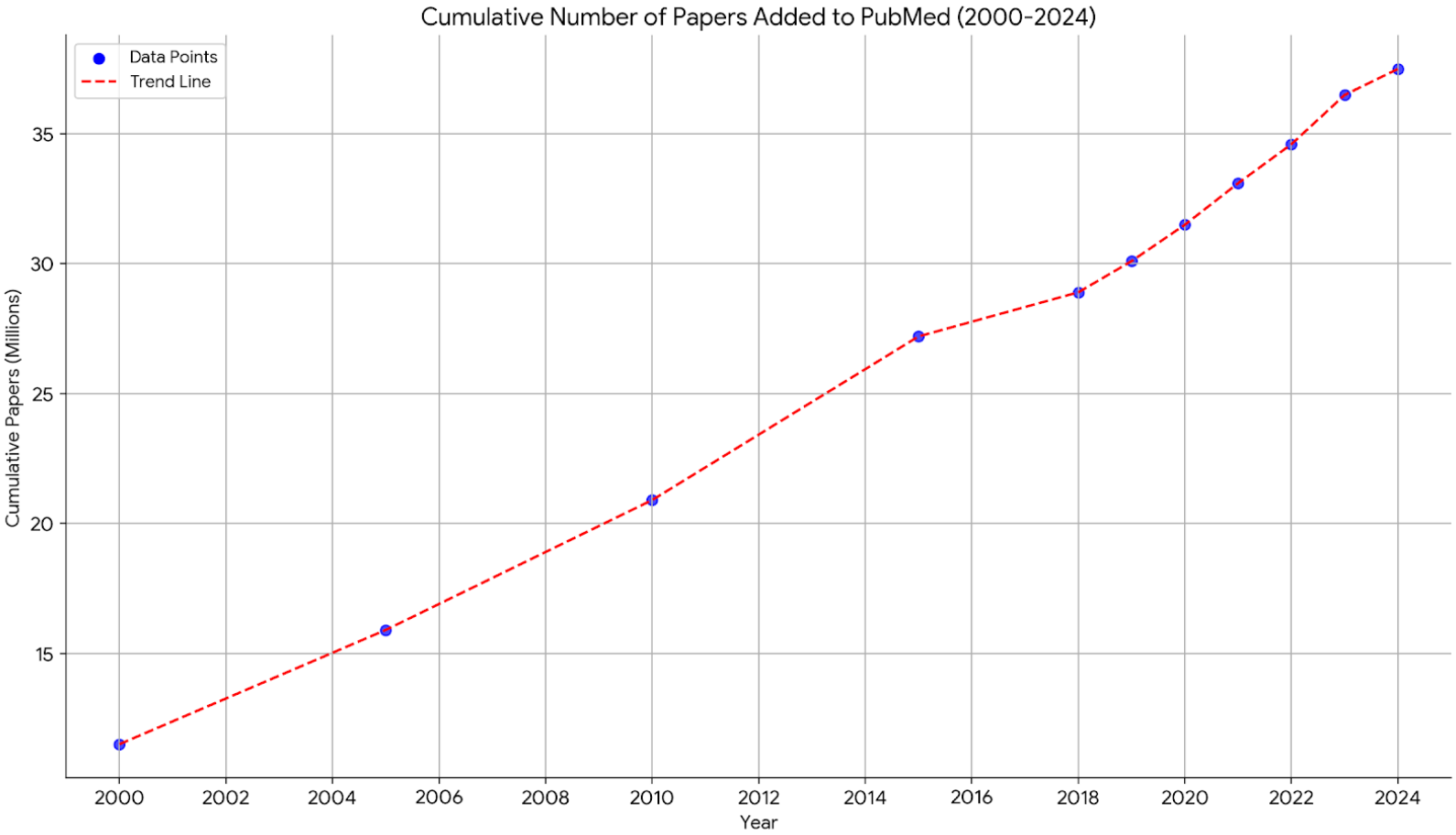
Exp growth of Data volumes & Computational capacity & Data storage per dollar



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.

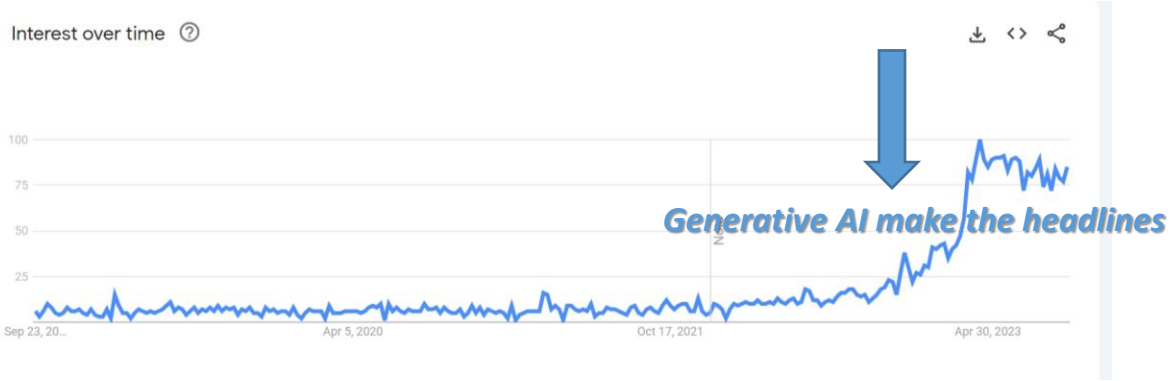
Category	Approx. rate now
All global data	**~7.0 PB/second**
Healthcare / biomedical proxy at 30%	**~2.1 PB/second**
Healthcare / biomedical proxy per year	**~66 ZB/year**

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.

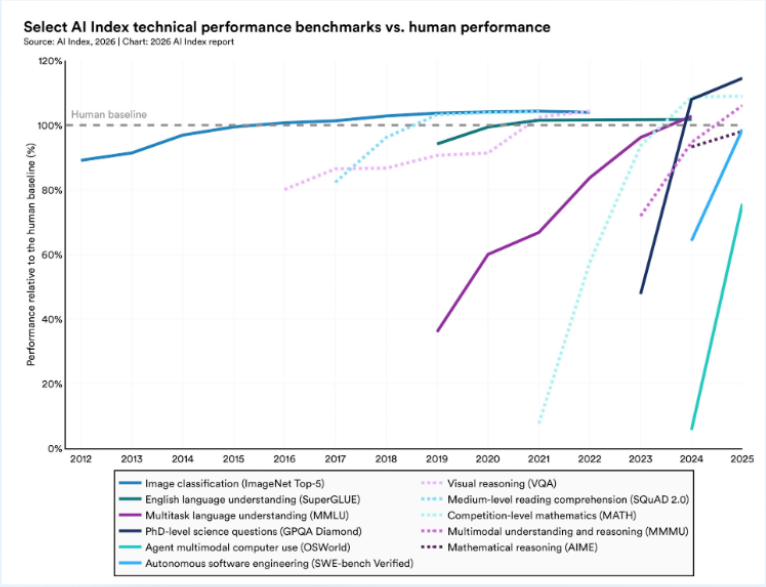
Machines are excellent at scale, speed, consistency, pattern detection, simulation, monitoring, and handling repetitive tasks. Humans remain essential for judgment, context, ethics, empathy, creativity, accountability, and deciding what outcomes actually matter.

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.

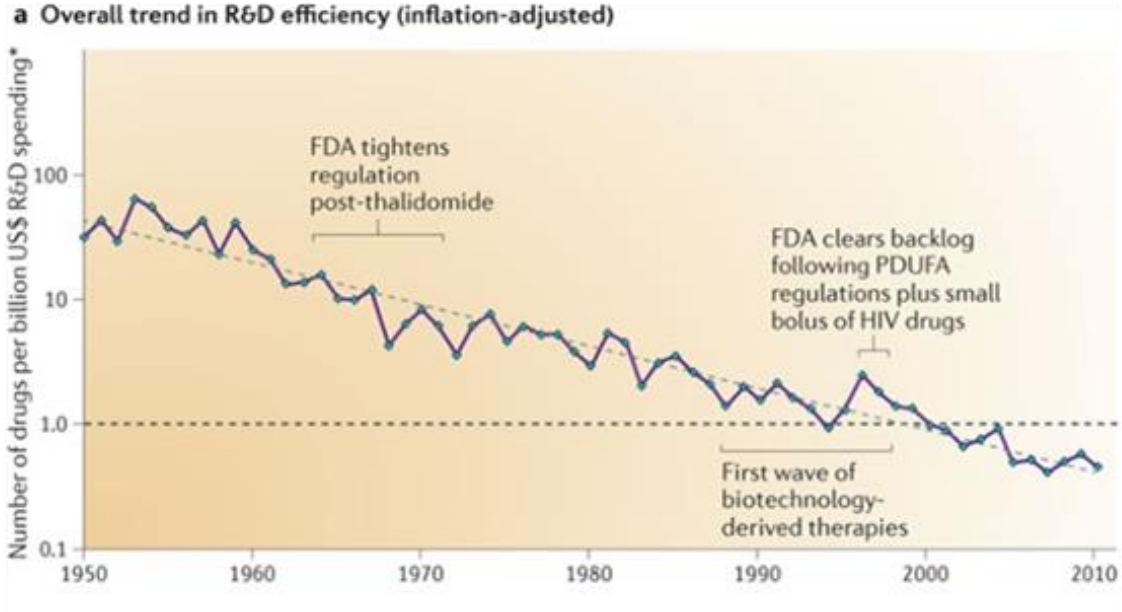
See: Chapter 2: Technical Performance



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



ELSEVIER

Drug Discovery Today

Volume 29, Issue 11, November 2024, 104160

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

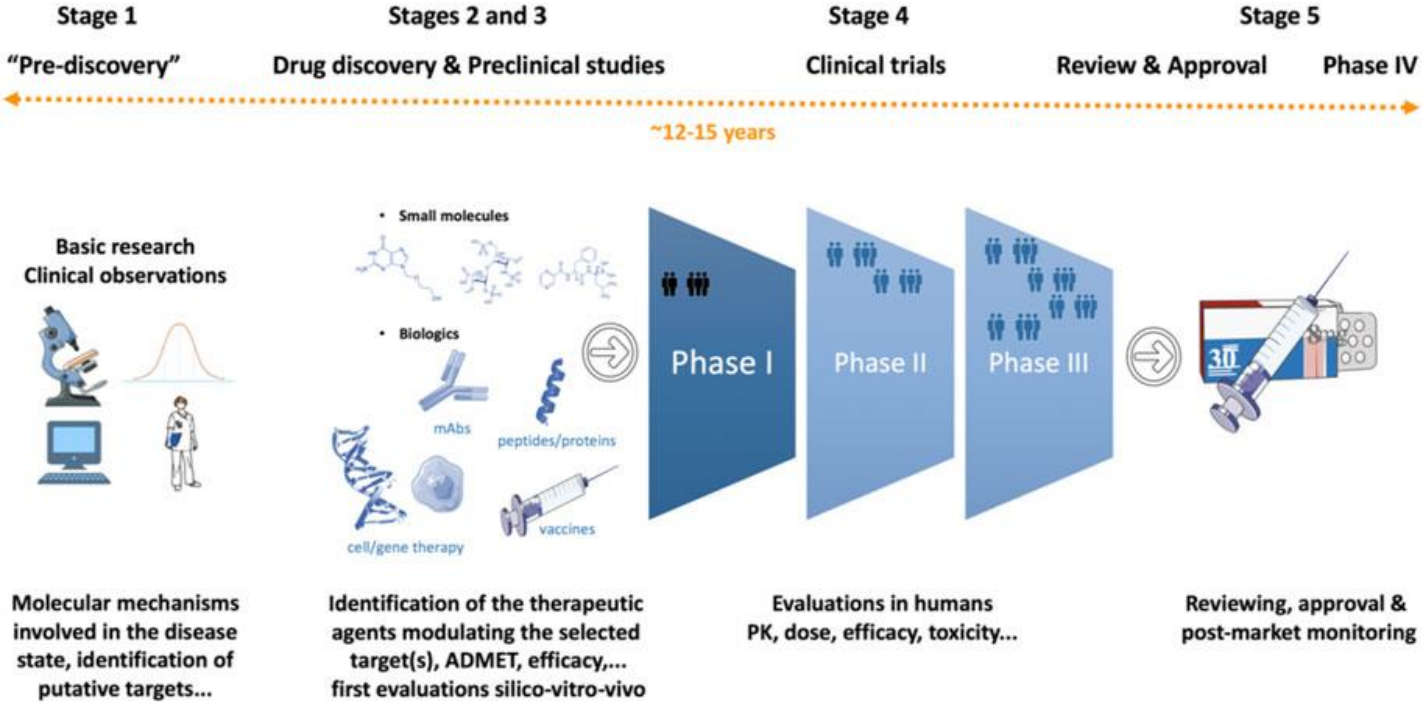
https://doi.org/10.1016/j.drudis.2024.104160 Get rights and content

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



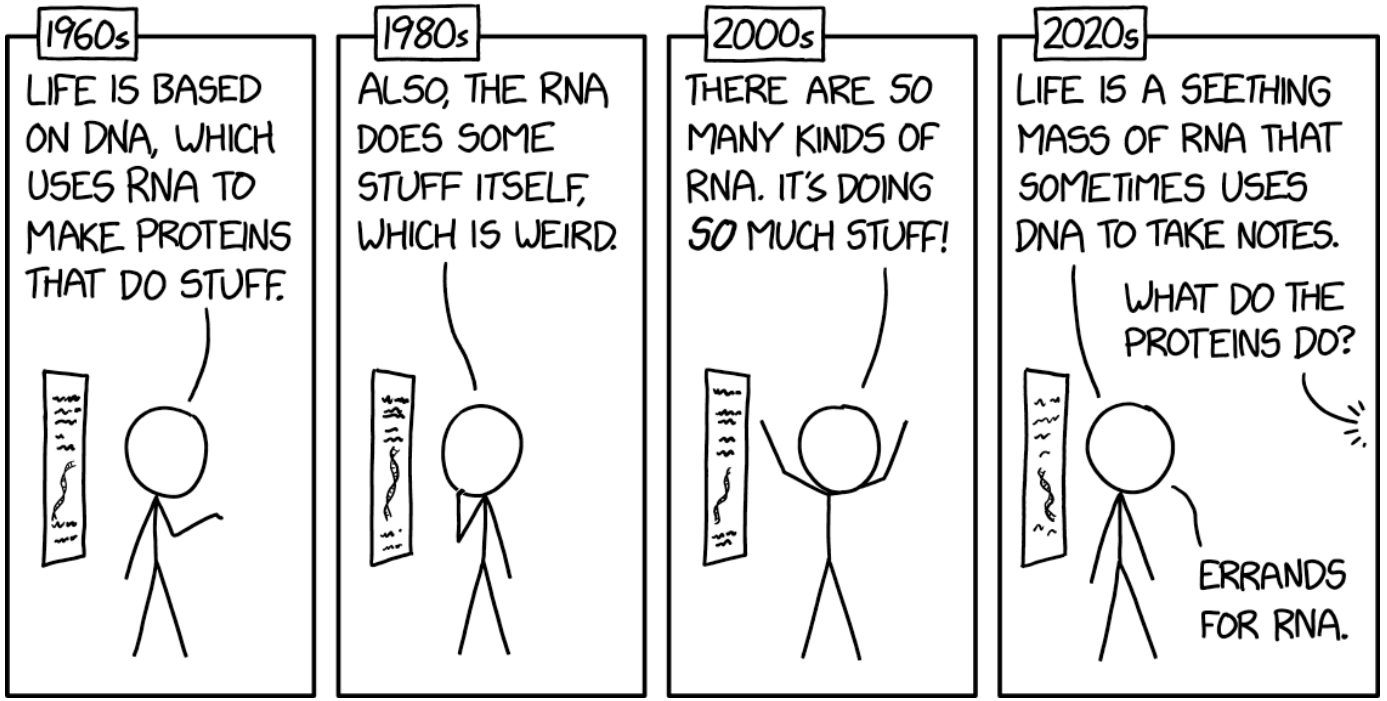
https://www.frontiersin.org/journals/drug-discovery/articles/10.3389/fddsv.2023.1201419/full

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

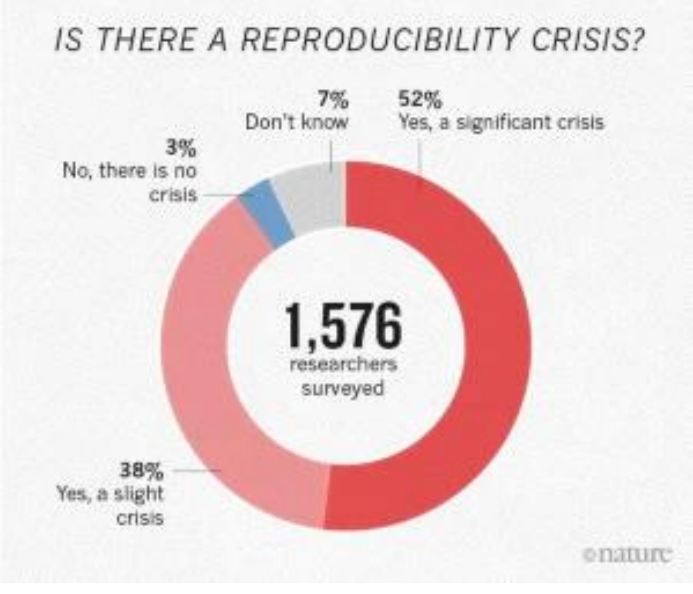
BIOLOGY is complex and understanding takes time

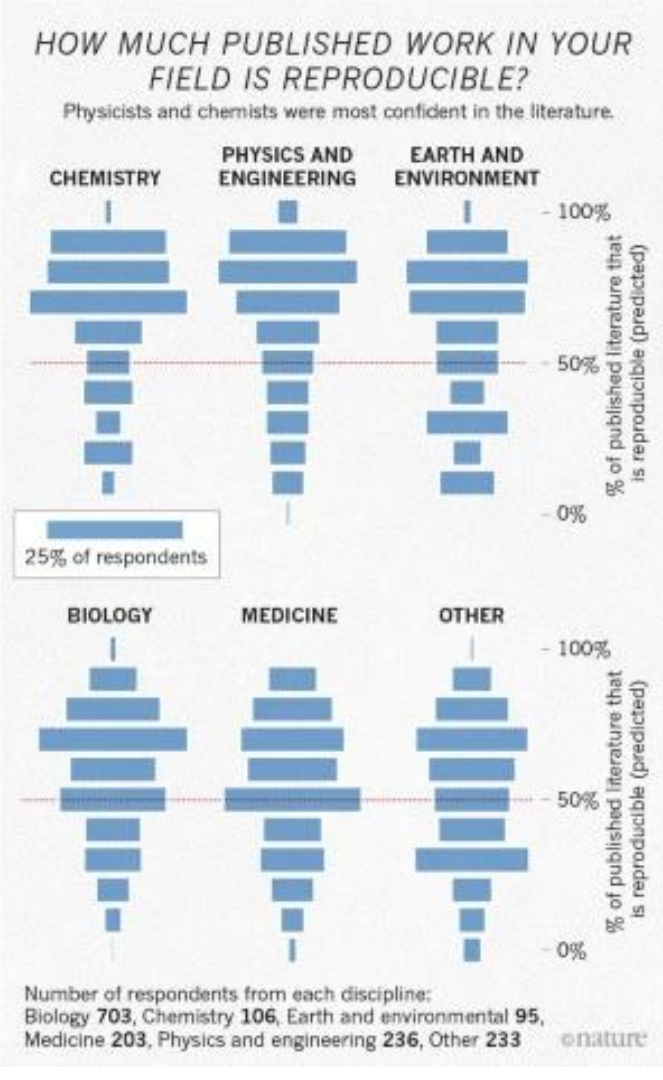


https://xkcd.com/3056/

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

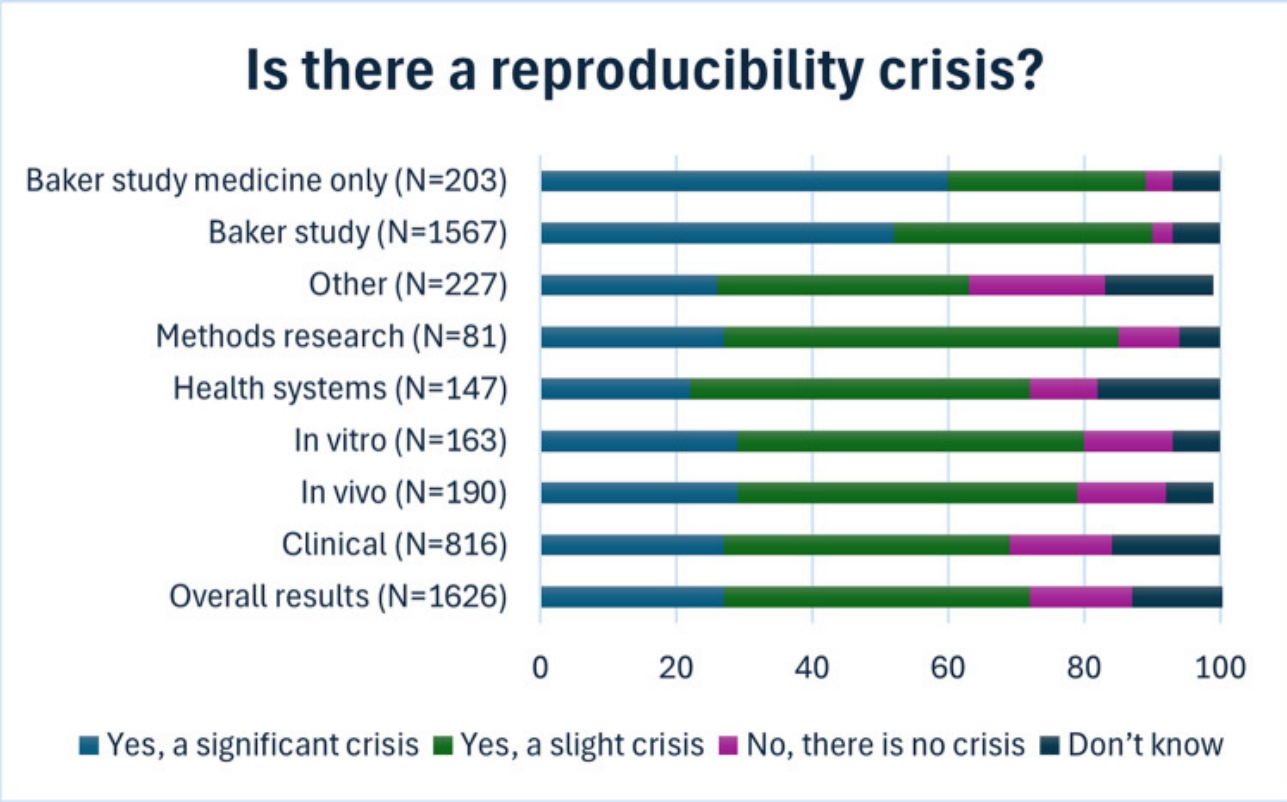




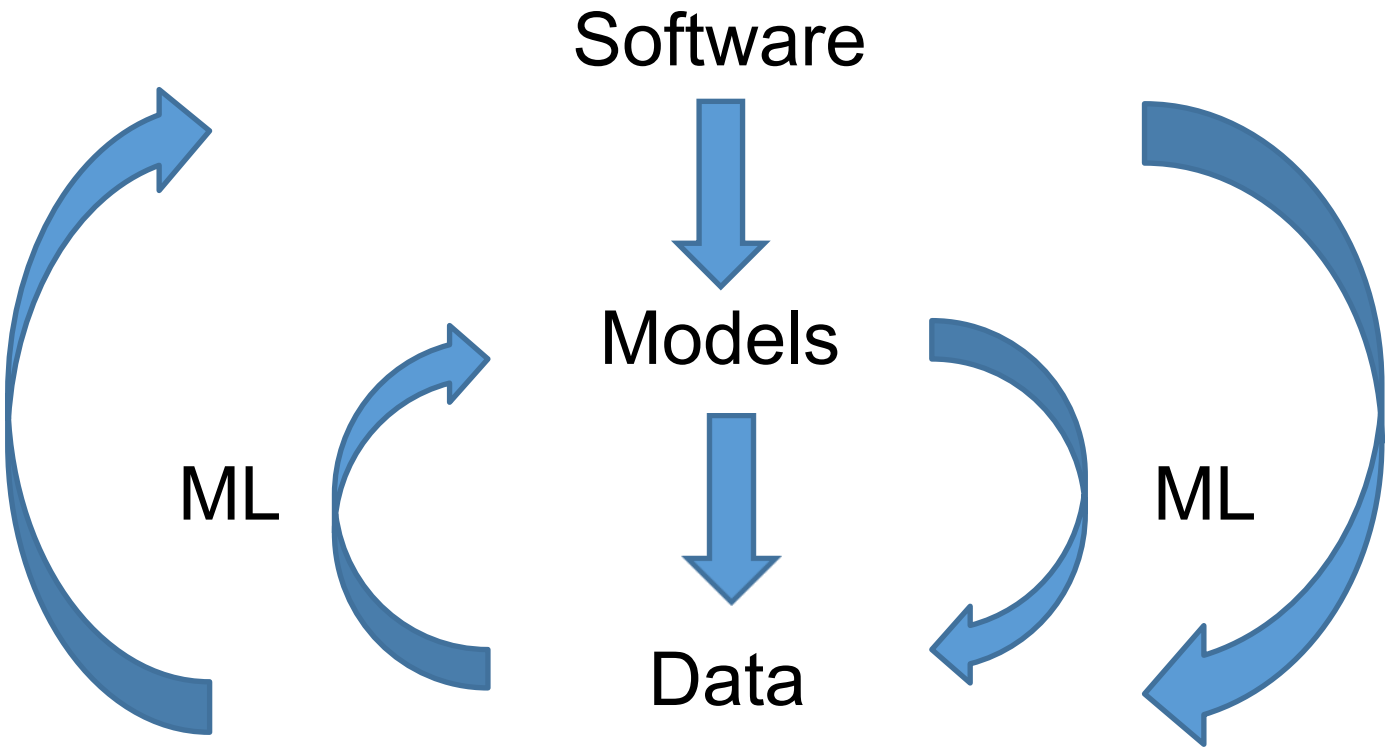
"We report the results of an international survey examining perceptions of reproducibility. Almost 3 quarters of participants reported that they felt there was a reproducibility crisis in biomedicine. The concern appears to apply to biomedicine overall, but also specifically to clinical research, in 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

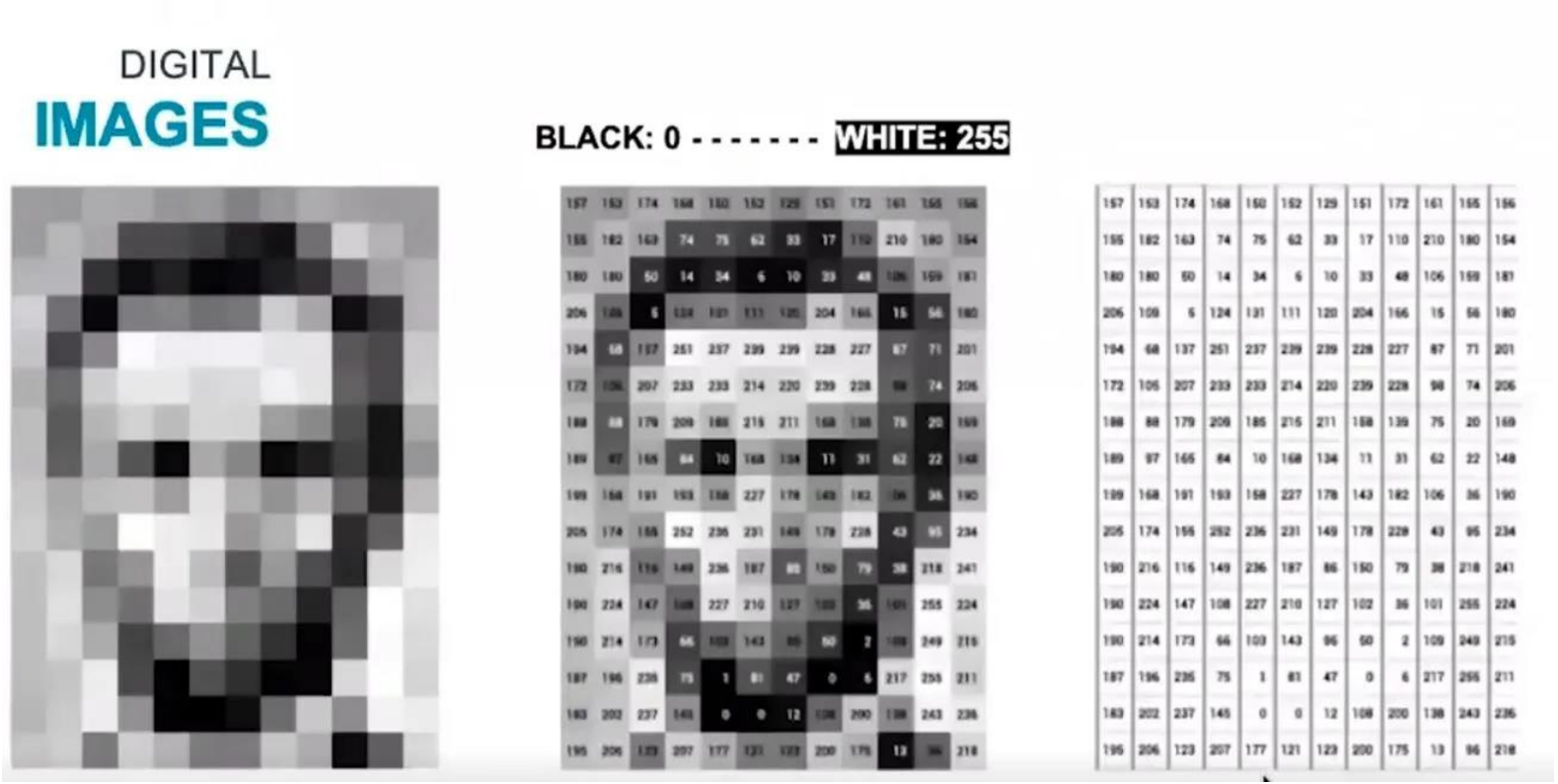


What is AI ?
Traditional vs AI- Machine Learning Computational Analysis



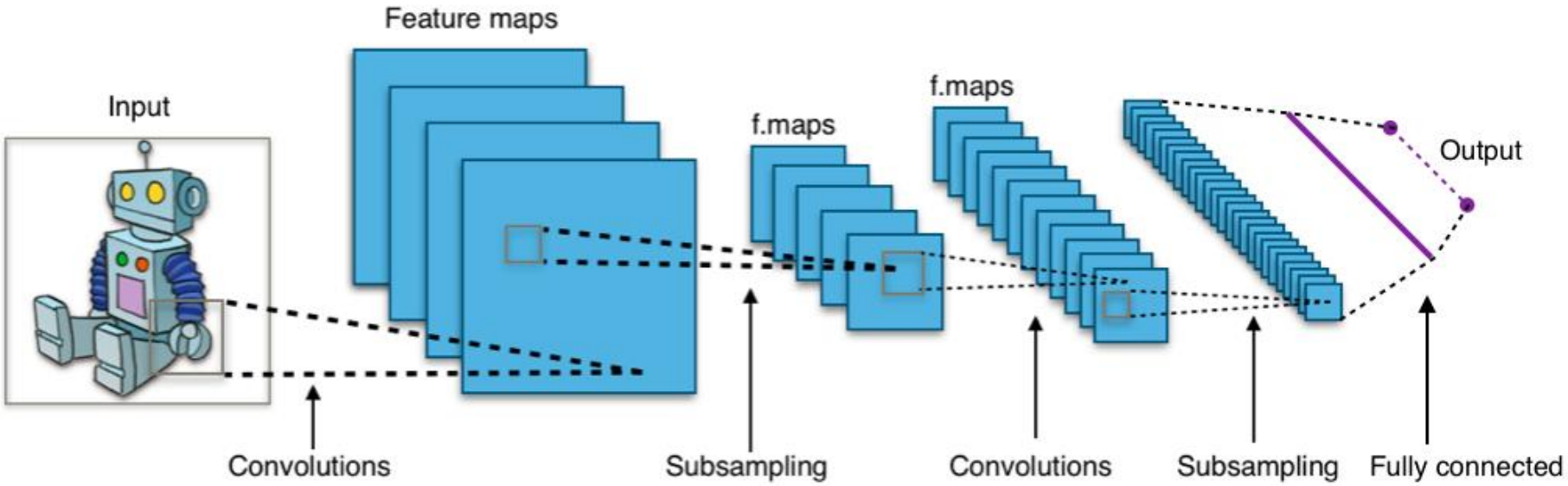
Deep Learning (Convolutional Neural Networks)) was first successful in Image Analysis

Transformer : Translating light intensities in pixels to numbers



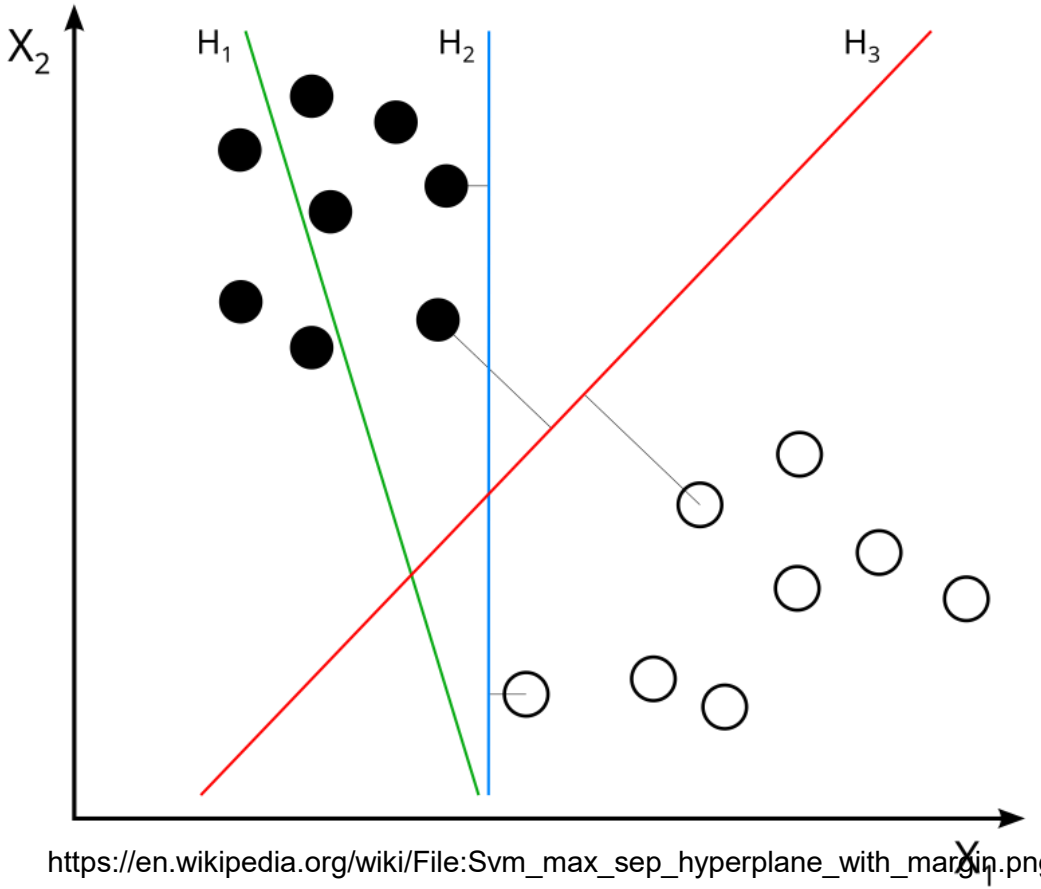
https://medium.com/@md-jewel/understanding-digital-images-for-image-processing-and-computer-vision-part-1-cc42be78cca1

Expanding the variable (feature) space



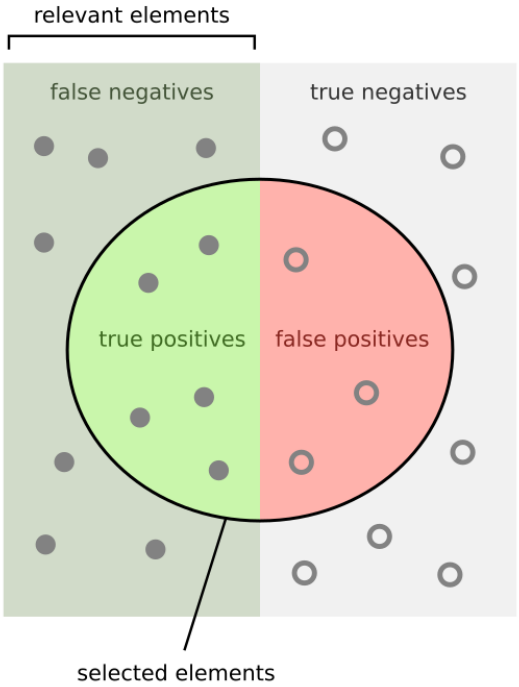
https://commons.wikimedia.org/wiki/File:Typical_cnn.png

And apply traditional Machine Learning like Support Vector Machines (SVMs)
An example of how to design a Predictor



https://en.wikipedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png

Basic characteristics for Predictors



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

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{relevant elements}}$$

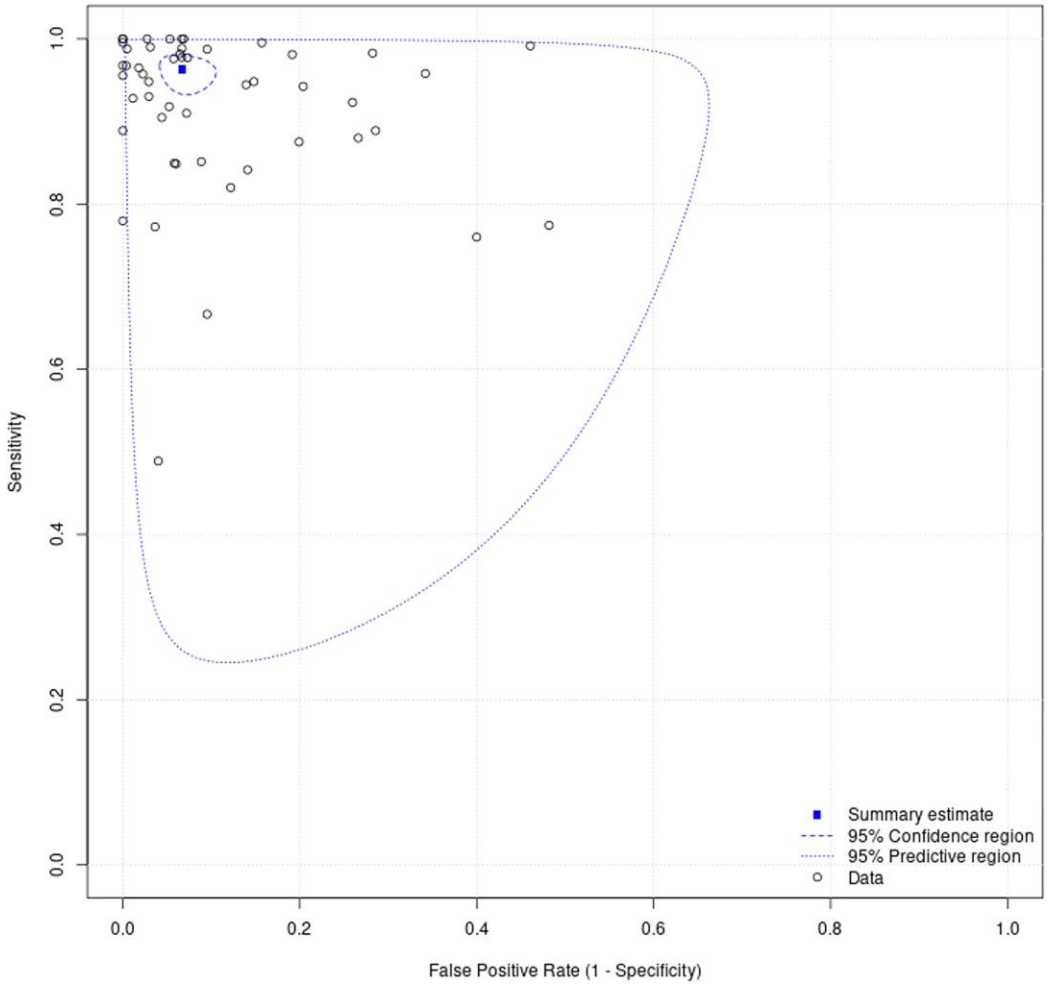
How many negative selected elements are truly negative? e.g. How many healthy people are identified as not having the condition.

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

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

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 Cartlidge, Henschel Freduah-Agyemang, Deborah D. Stocken & Darren Treanor
npj Digital Medicine volume 7, Article number: 114 (2024) Cite this article

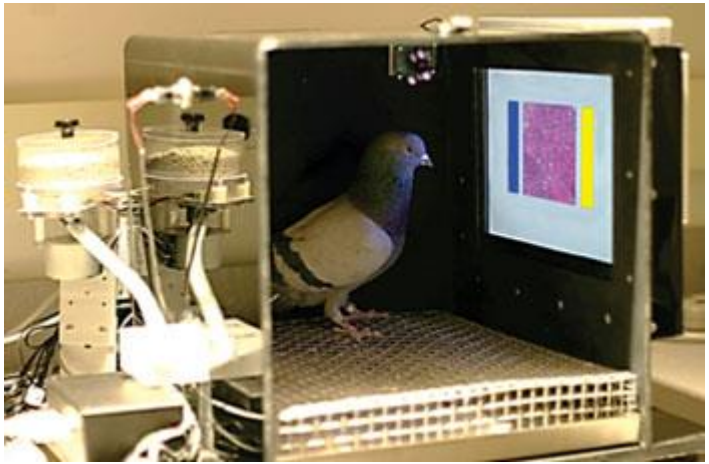
The Human Baseline

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

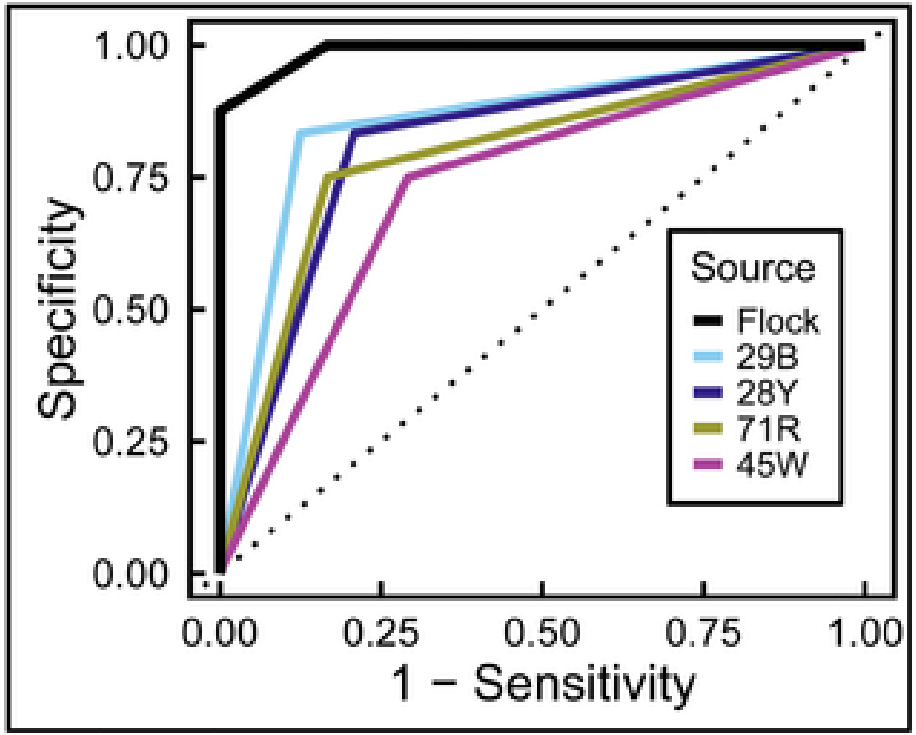
Consensus Diagnosis	Mean (95%CI) Concordance Rate	Above Consensus Diagnosis (95% CI)	Below Consensus Diagnosis (95% CI)	Mean (95%CI) Under/Over-calling Rate	Mean (95%CI) Review Time (sec)
Benign	71% (61%, 82%)	29% (20%, 40%)	-	.29 (.18, .40)	120.6 (101.6, 139.5)
Atypia	37% (29%, 45%)	21% (15%, 28%)	43% (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)

Brunyé TT, Mercan E, Weaver DL, Elmore JG. Accuracy is in the eyes of the pathologist: The visual interpretive process and diagnostic accuracy with digital whole slide images. J Biomed Inform. 2017 Feb;66:171-179. doi: 10.1016/j.jbi.2017.01.004. Epub 2017 Jan 10. PMID: 28087402; PMCID: PMC5316368.

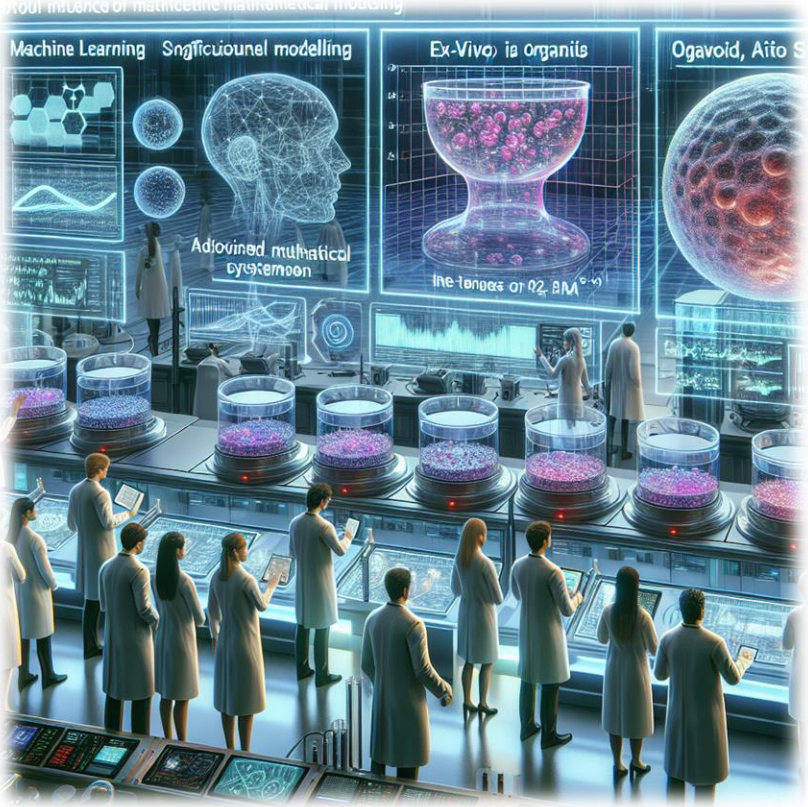
And how good are Pigeons?



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



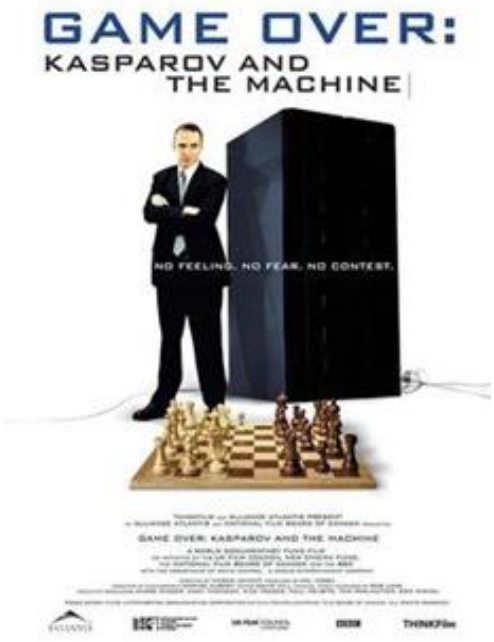
It is hard to predict, especially the future



“A chess problem is genuine mathematics, but it is in some way 'trivial' mathematics”. From A Mathematician's Apology, a1940 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:

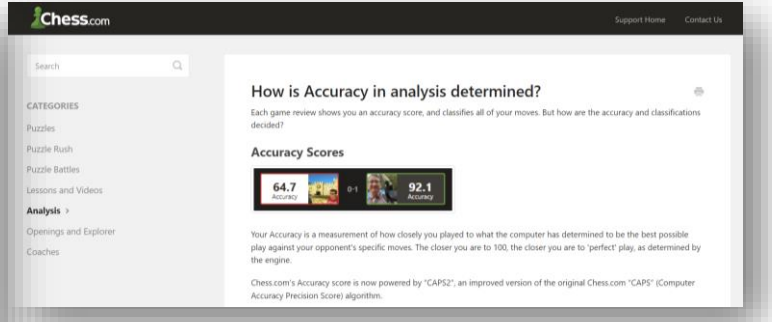
"A weak human + machine + better process was superior to a strong computer alone and … superior to a strong human + machine + inferior process."

"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…



- ***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**

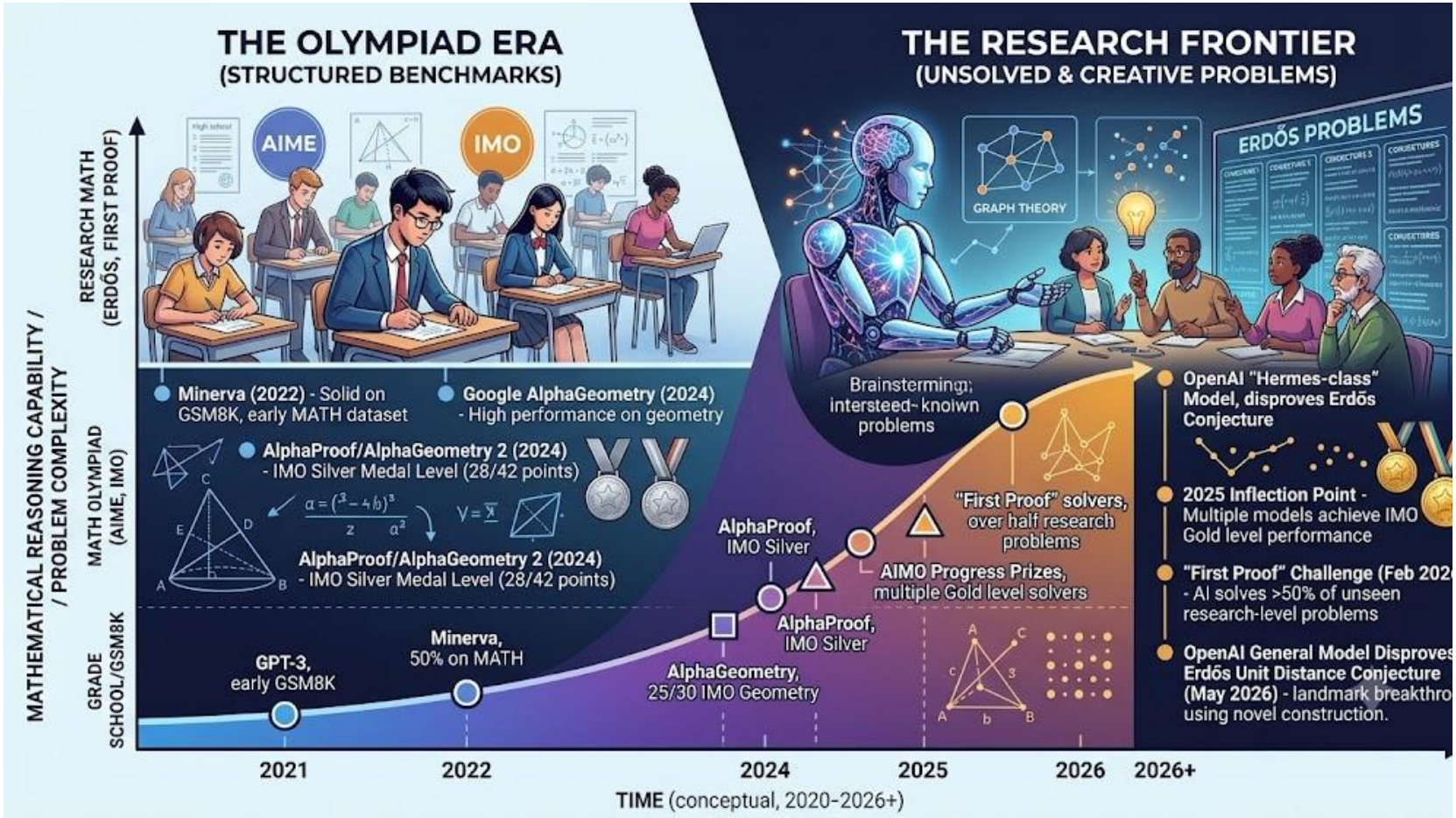
Hikaru Nakamura + Magnus Carlsen loose 100% of their chess games *vs Stockfish*

https://www.youtube.com/watch?v=0T-1sG-sZLU

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 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 Gehrunger, 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

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 expositionally refined through human interactions with Codex.

DECLARATION TEXT

Leiden Declaration on Artificial Intelligence and Mathematics

2 June 2026 · DOI: 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 EFFECTIVENSS OF MATHEMATICS IN THE NATURAL SCIENCES, Eugene Wigner, Communications in Pure and Applied Mathematics, Vol. 13, No. I (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

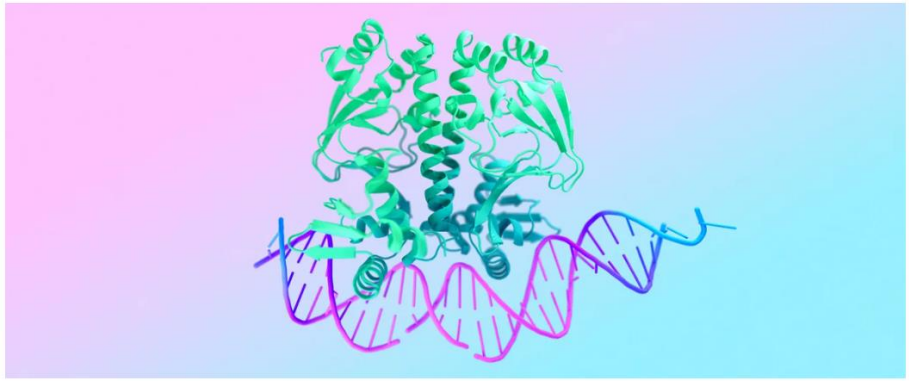
May 08, 2024
6 min read

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

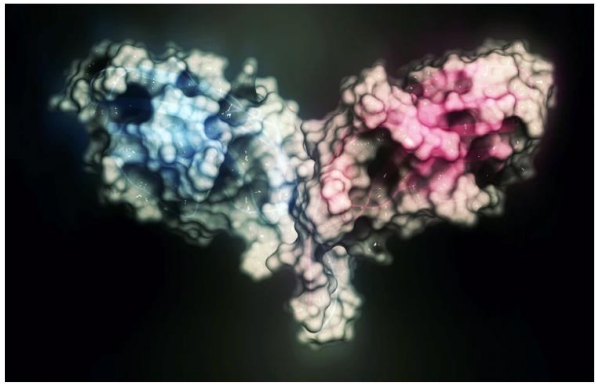
nature › news › article

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

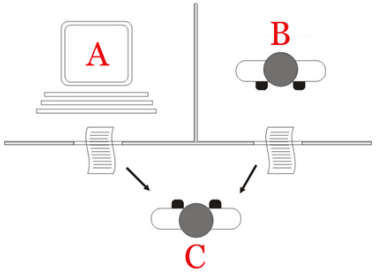
A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

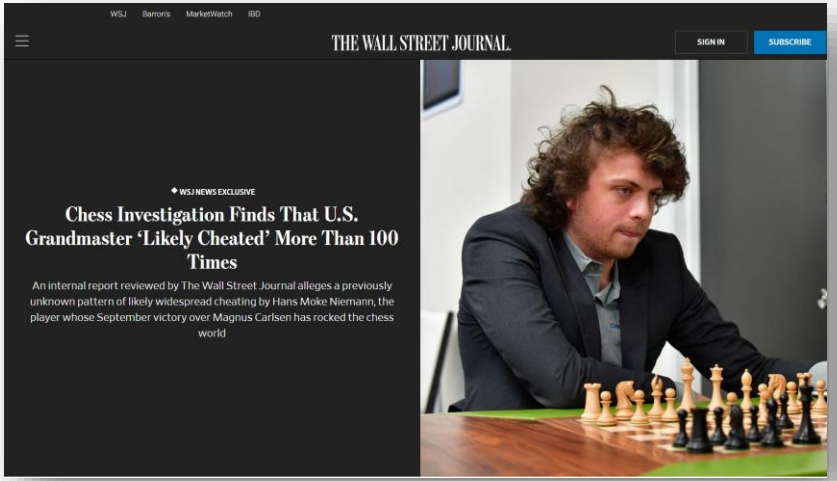
By A. M. Turing

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…

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
COMPUTER SCIENCES OPEN ACCESS

Large language models pass a standard three-party Turing test

Cameron R. Jones[a,b,1] and Benjamin K. Bergen[a]

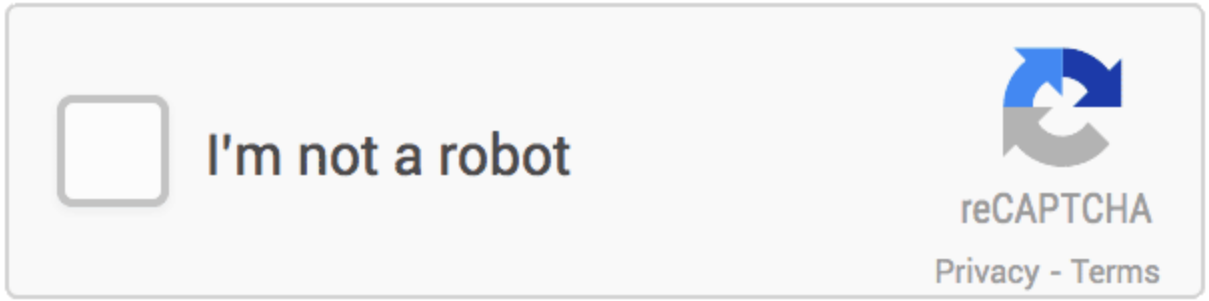
Edited by Brenden Manker Lake, Princeton University, Princeton, NJ; received September 16, 2025; accepted March 27, 2026 by Editorial Board Member Michael S. Gazzaniga

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)

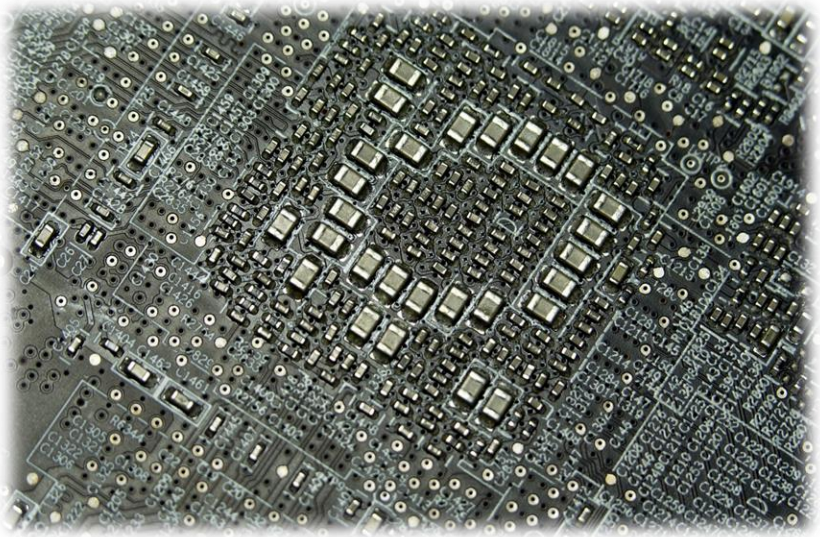


By Google - https://www.google.com/recaptcha/intro/index.html, Fair use, https://en.wikipedia.org/w/index.php?curid=45398292

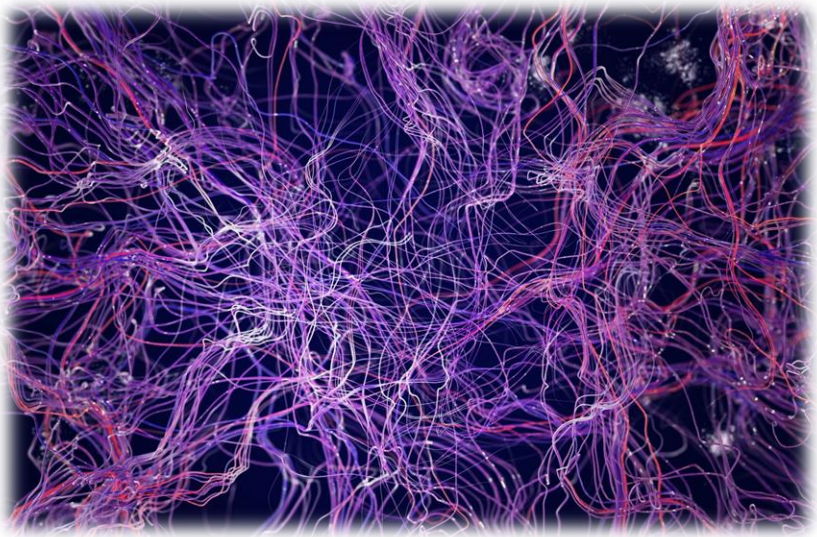
Basic underlying test principles:

- 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



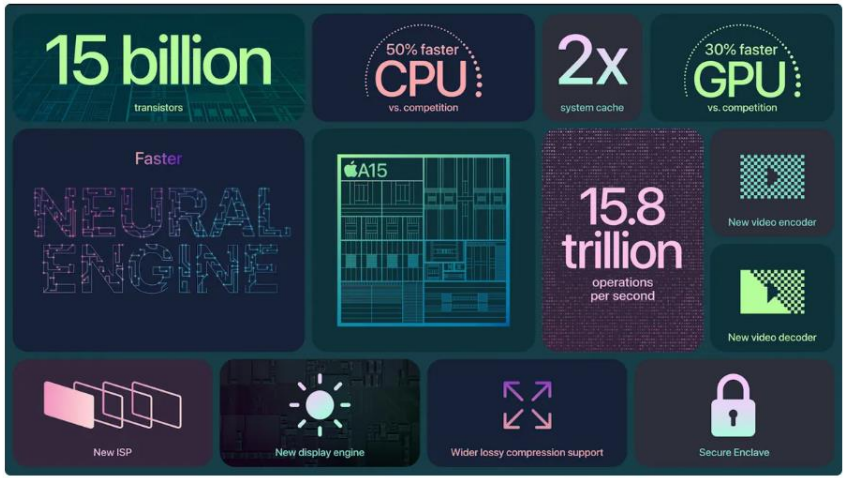
https://thewire.in/science/new-transistor-design-combines-best-high-low-performance-versions



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

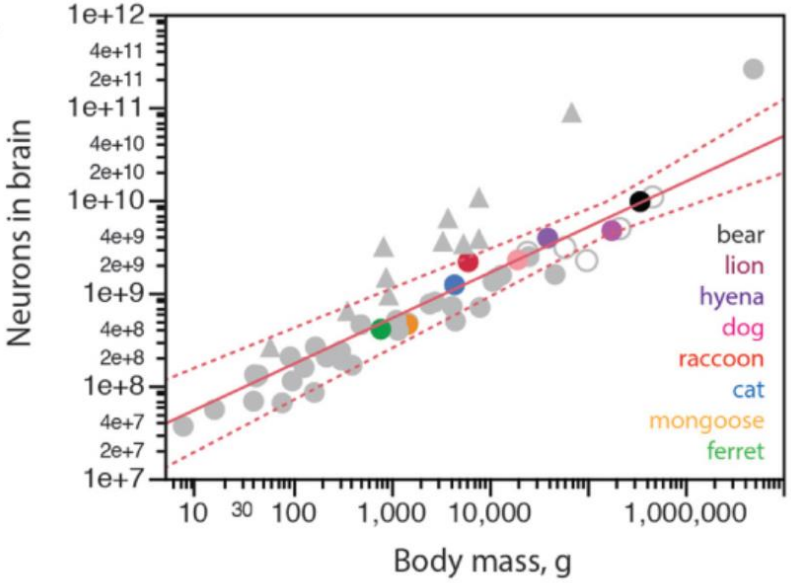


Apple A15 Bionic processor details

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

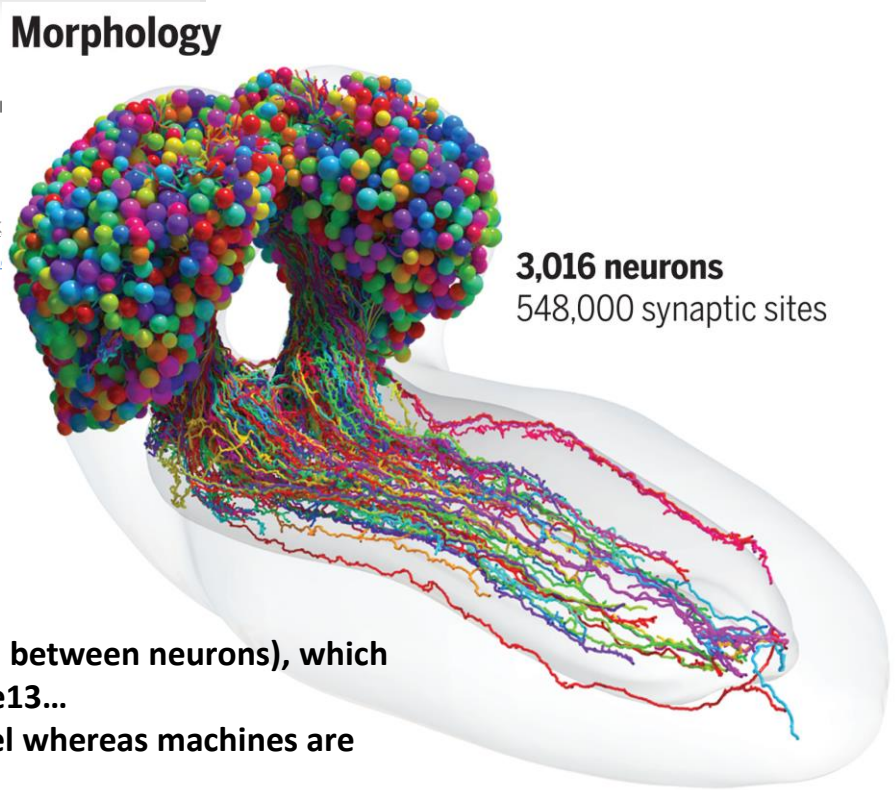
RESEARCH ARTICLE | NEUROSCIENCE

The connectome of an insect brain

MICHAEL WINDING, BENJAMIN D. PEDIGO, CHRISTOPHER L. BARNES, HEATHER G. PATSOLIC, YOUNGSER PARK, AKIRA FUSHIKI, INGRID V. ANDRADE, AVINASH KHANDELWAL, [...], 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**

nature methods

Editorial | Published: 08 December 2025

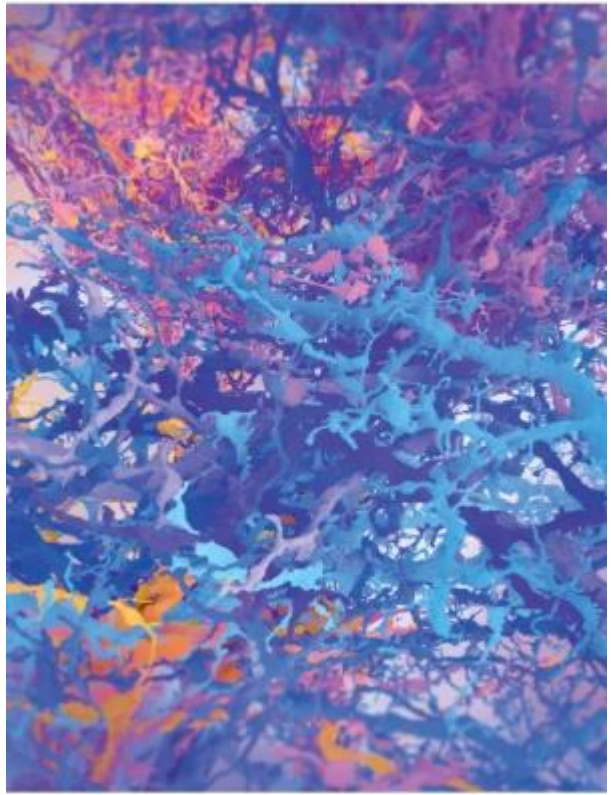
Method of the Year 2025: electron microscopy-based connectomics

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

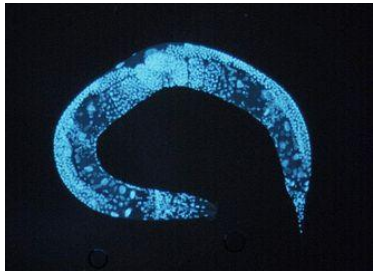
Published: 09 January 2008

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

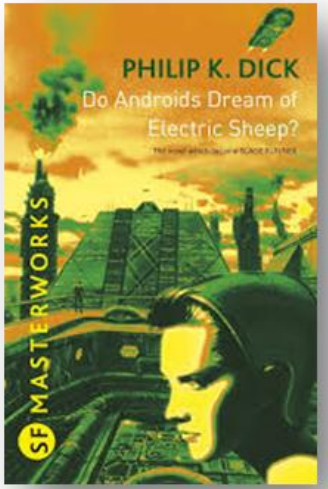
A Corrigendum to this article was published on 12 June 2008



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 moults[4]. Like sleep,

Natural Intelligence -- Animals dream



PNAS BRIEF REPORT | EVOLUTION OPEN ACCESS Check for updates

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

Cold Spring Harbor Laboratory

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

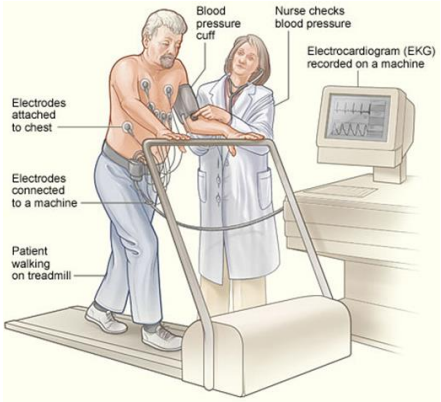
doi: https://doi.org/10.1101/2023.08.03.551880

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


nature

Explore content | About the journal | Publish with us | Subscribe

nature > news > article

NEWS | 18 January 2023

ChatGPT listed as author on research papers: many scientists disapprove

At least four articles credit the AI tool as a co-author, as publishers scramble to regulate its use.

Chris Stokel-Walker

?

- 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

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.

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

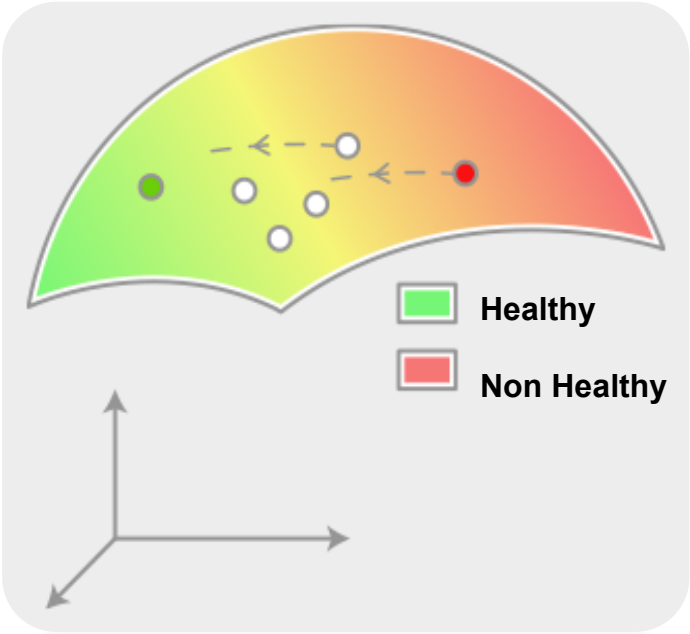
The Holy grail of modern biomedical research is to *understand and control* the human immune system

The immune system in

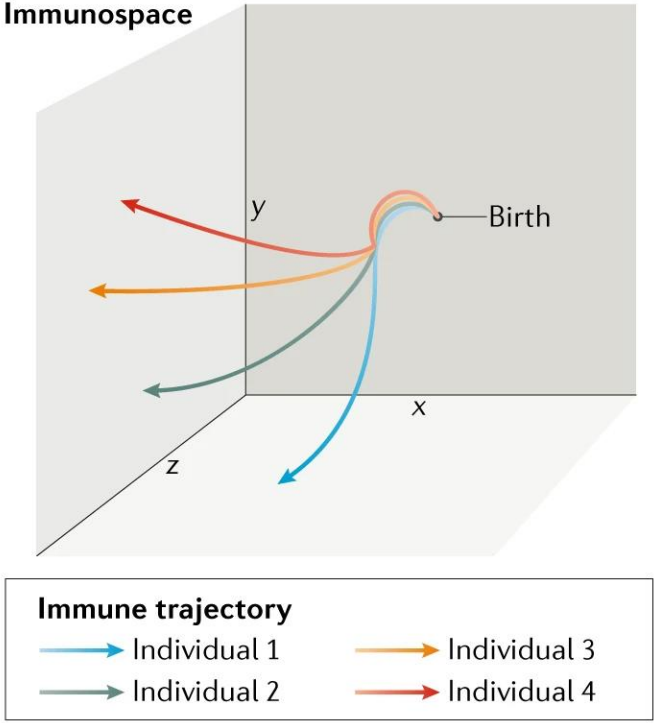
- **Autoimmune disease**
- **Cancer**
- **Infectious disease**
- **Neurodegenerative disease**
- **And most other human disease conditions…**

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**

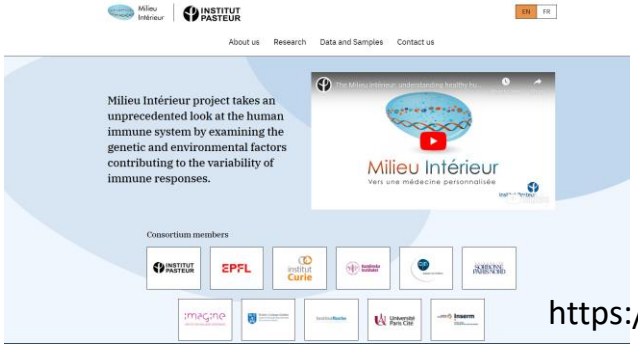


General Systems Immunology knowledge suggest e.g. using individuals as their own control



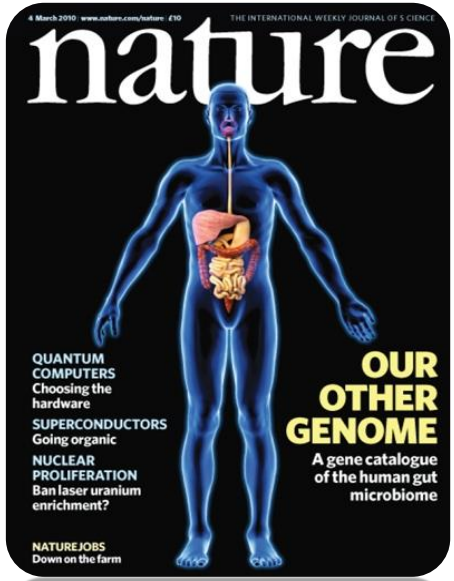
- "Human immune systems are relatively stable within individuals over the course of weeks to months, but incredibly variable between individuals"
- "…induced responses to pathogens differ markedly among different age groups and… these differences are unique to different kinds of stimuli"

Petter Brodin
Nature Reviews | Immunology volume 19 | February 2019 | 87



https://www.milieuinterieur.fr/en/

The human model is complex and context dependent



Nature March 2010

Nature June 2012

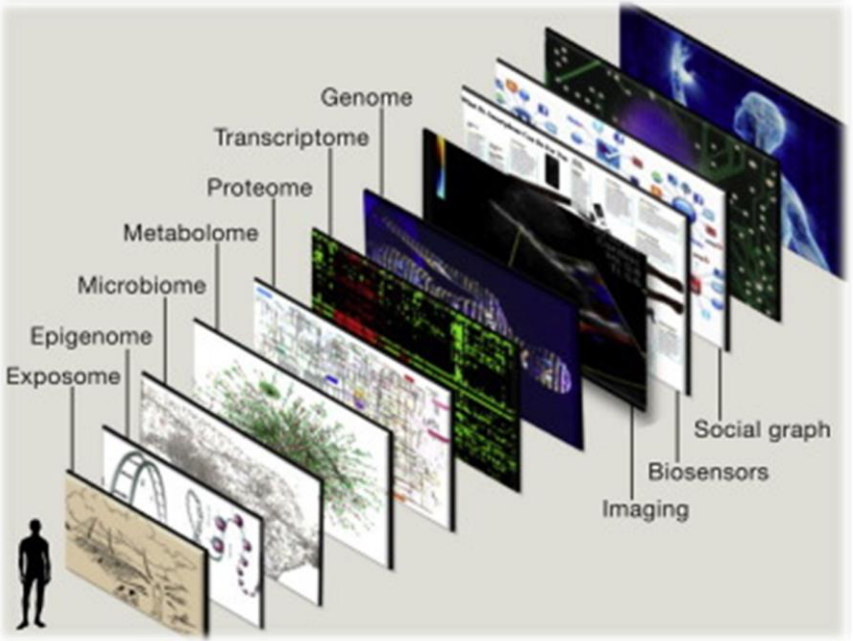
- 3 × 10^13 human cells
- As many bacterial cells
- Around 10 times as many viruses.

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 ←→**
- **Cells ←→**
- **Tissues←→**
- **Organs ←→**
- **Organ complexes ←→**
- **Organisms ← →**
- **Populations ← →**
- **….**

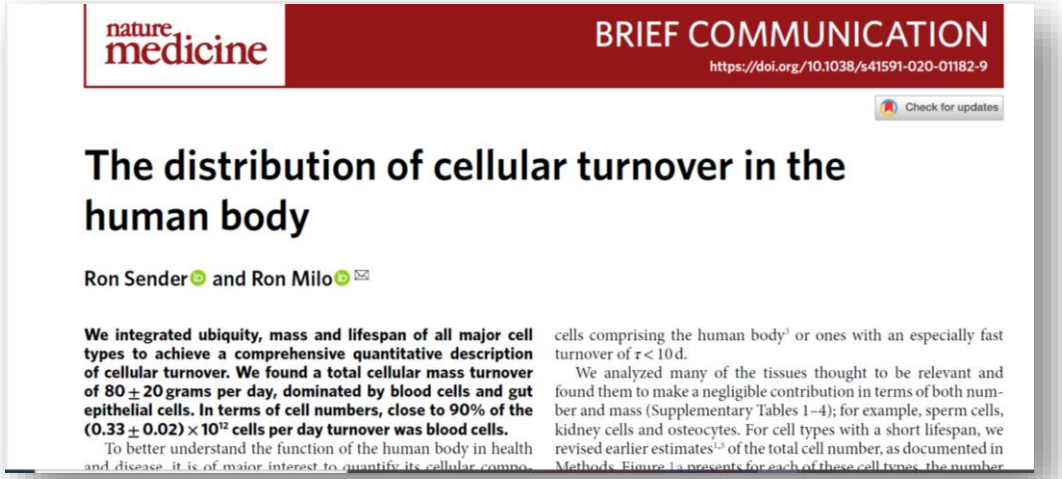



Leading Edge
Review
Cell

Individualized Medicine
from Prewomb to Tomb

Eric J. Topol[1,*]
[1]The Scripps Translational Science Institute, The Scripps Research Institute and Scripps Health, La Jolla, CA 92037, USA
*Correspondence: etopol@scripps.edu
http://dx.doi.org/10.1016/j.cell.2014.02.012

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

Check for updates

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 ± 0.02) × 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 body[1] or ones with an especially fast turnover of $\tau < 10$ d.

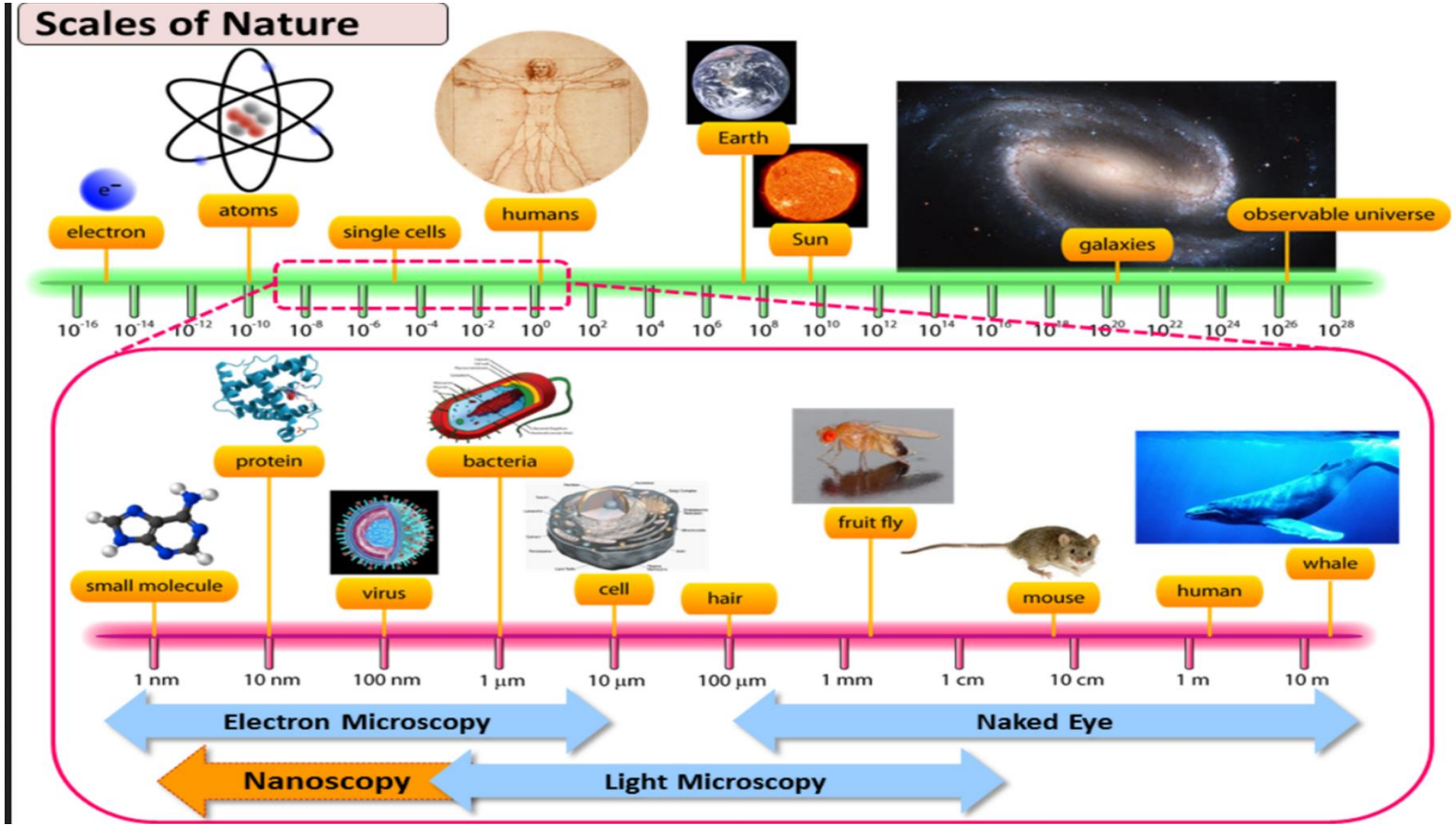
We analyzed many of the tissues thought to be relevant and found them to make a negligible contribution in terms of both number and mass (Supplementary Tables 1–4); for example, sperm cells, kidney cells and osteocytes. For cell types with a short lifespan, we revised earlier estimates[1,3] of the total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

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

"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 ± 0.02) × 10^12 cells per day turnover was blood cells."

"The most recent rigorous census has estimated the number of cells in the adult body to be approximately 30 ± 0.5 × 10^12, of which about 90% are from the hematopoietic lineage, mostly red blood cells."

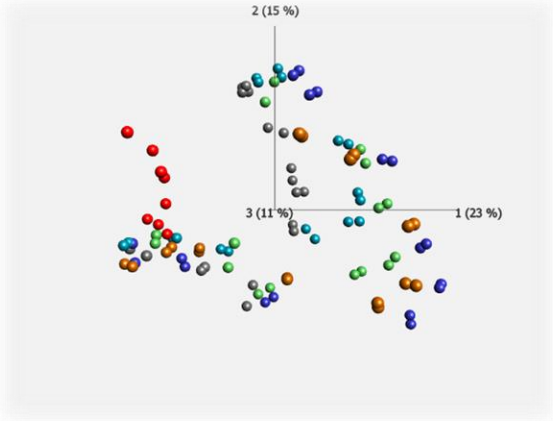
Some (log) length scales of Life

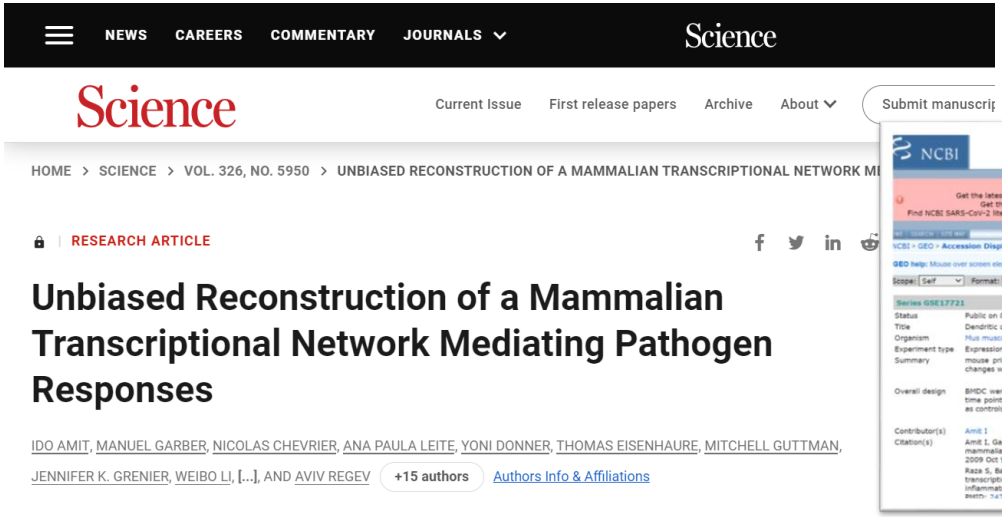


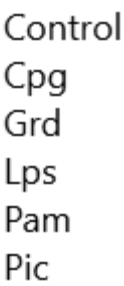
http://www.bates.edu/gould-research-lab

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:





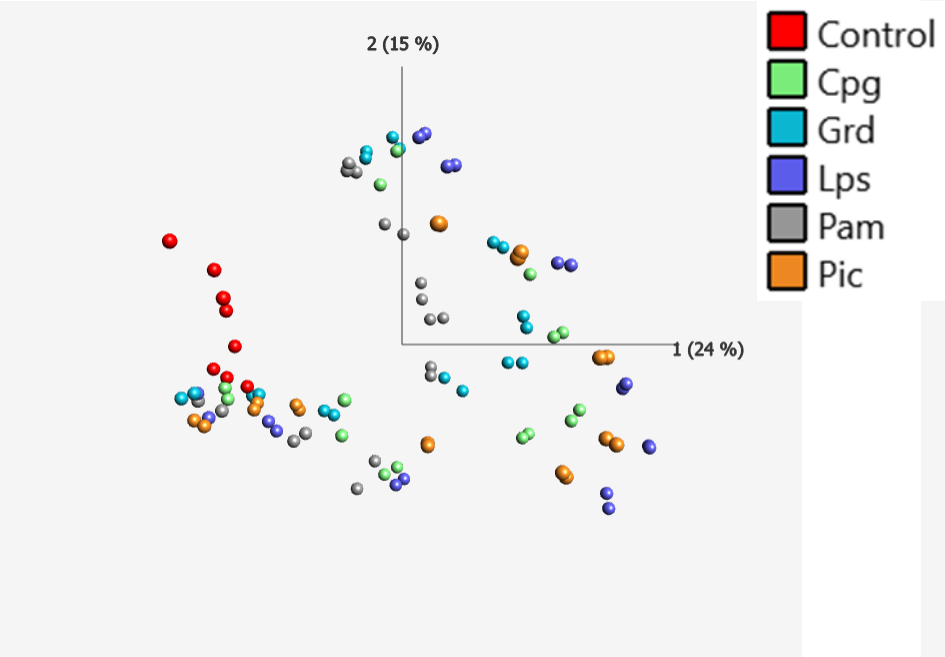


A small sub-part of the full 96 times 10716 matrix, representing 10716 genes and 96 samples for the invitro stimulated mouse dendritic cell data:

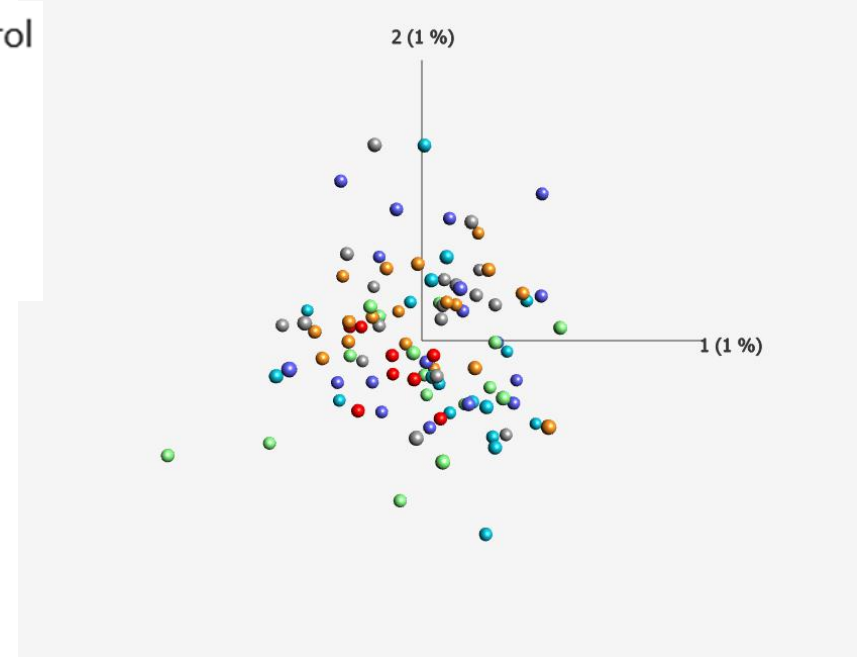
	p-values	q-values	ol	Control	Control	Control	Control	Control	Control	Control	Cpg	Cpg
Myo5a				374.6	381.35	408.6	387.6	424.55	411.5	396.1	376.25	360.65
Nabp1			5	685.6	711.7	659.55	718.5	758.75	880.05	758.95	764.2	761.05
Nacc2				66.4	36.7	44.5	43.6	39.5	41.8	44.3	38.6	40
Myo1c				157.7	189.85	133.9	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
Nap1l1///...				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 anayze 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)

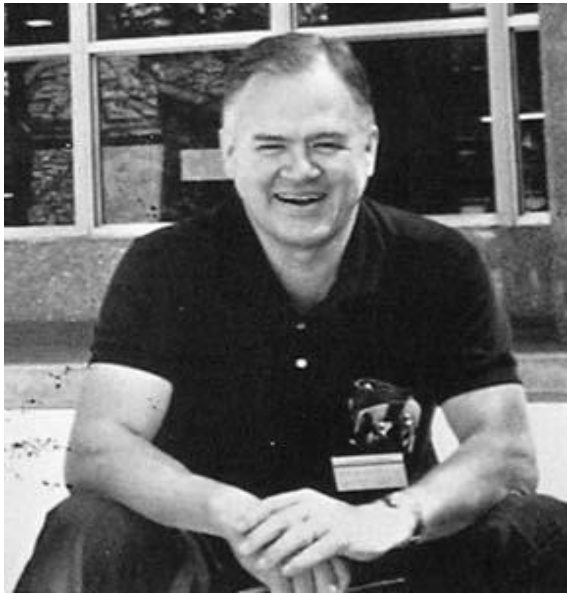


PCA on corresponding randomized data capturing 2% of the variance in 2D (3% in 3D)

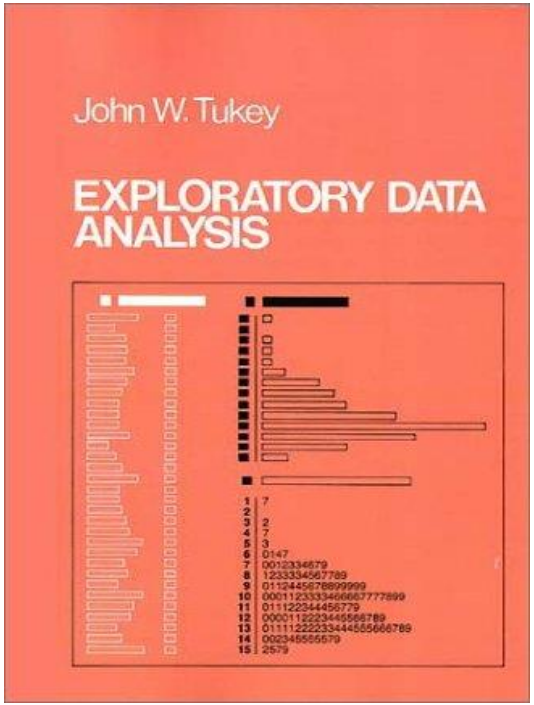
STATISTICAL LEARNING & VISUALIZATION → ACQUIRING INSIGHTS AROUND PATTERNS IN DATA RELATED TO BIOLOGICAL VARIATION

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

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

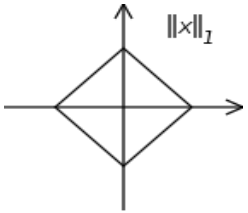
$$d(x, y) = d(y, x) \quad (\textit{symmetry}) \tag{2.2}$$

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

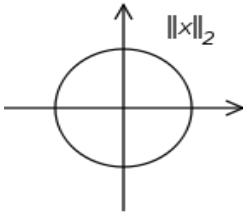
Examples:

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

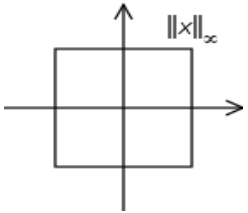
Effects of different similarity measures or distance functions



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).

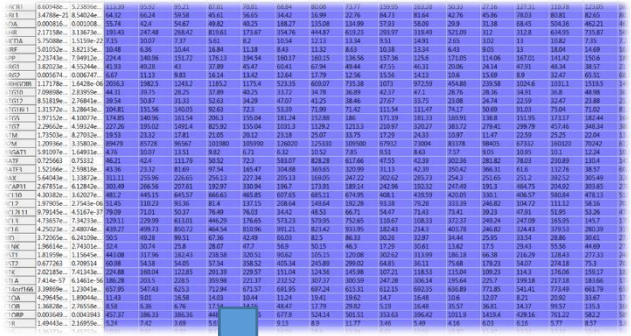


The normal “unit sphere” in the standard L^2 norm



The “unit sphere” in the L^∞ norm

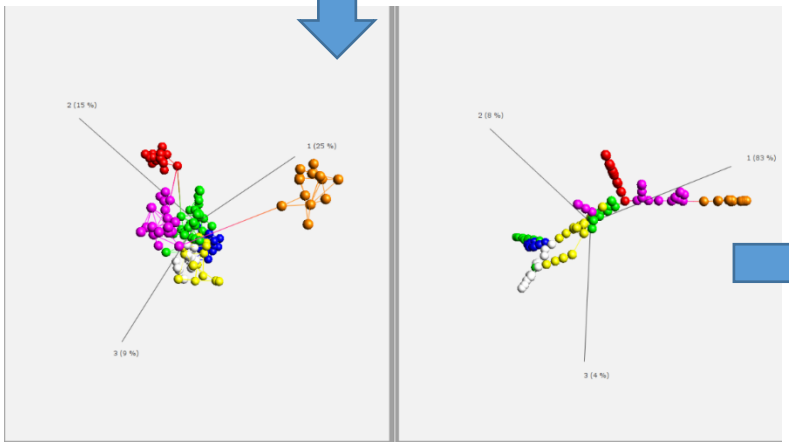
Feature extraction and selection to "understand" phenotypic variation



Measurements

Sample space population

Data



Data Analysis

→ clustering & classification → Biological Insights

How to explore MULTI-OMICS data?

- N=10-10^5 Samples
- P=10-10^10 Variables

How to find relevant structure?

Problems:

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

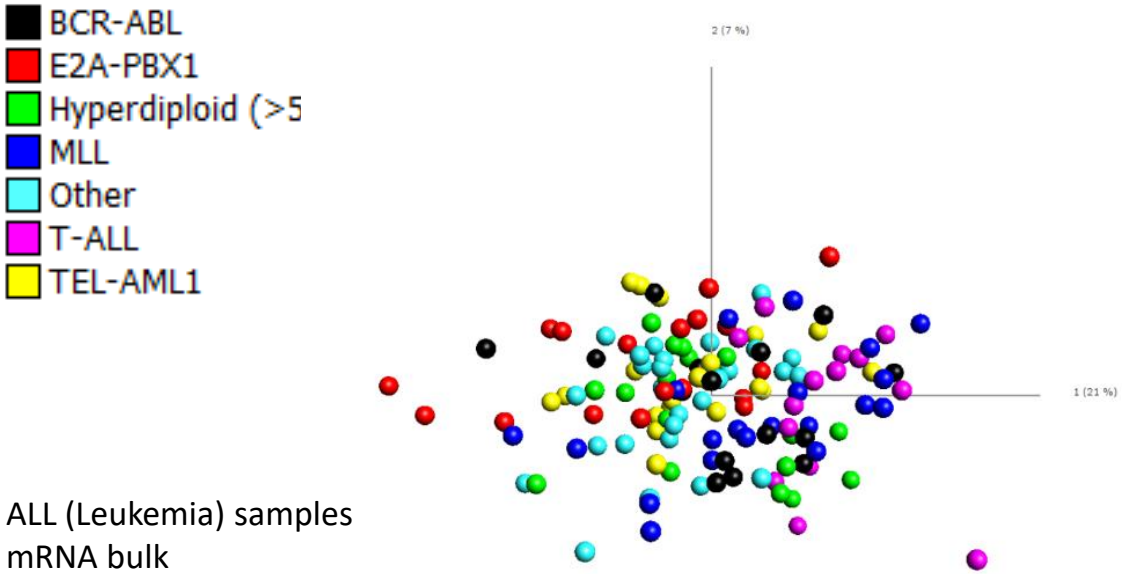
Visualize and Analyze the data

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

Classification of pediatric acute lymphoblastic leukemia by gene expression profiling
Mary E. Ross et al.
Blood 2003 102:2951-2959; doi:10.1182/blood-2003-01-0338
Data available at https://www.stjuderesearch.org/site/data/ALL3/

Example Data: Affymetrix chip 22282 ProbeIDs and 132 Samples

Dimension reduction techniques: *Classical PCA*

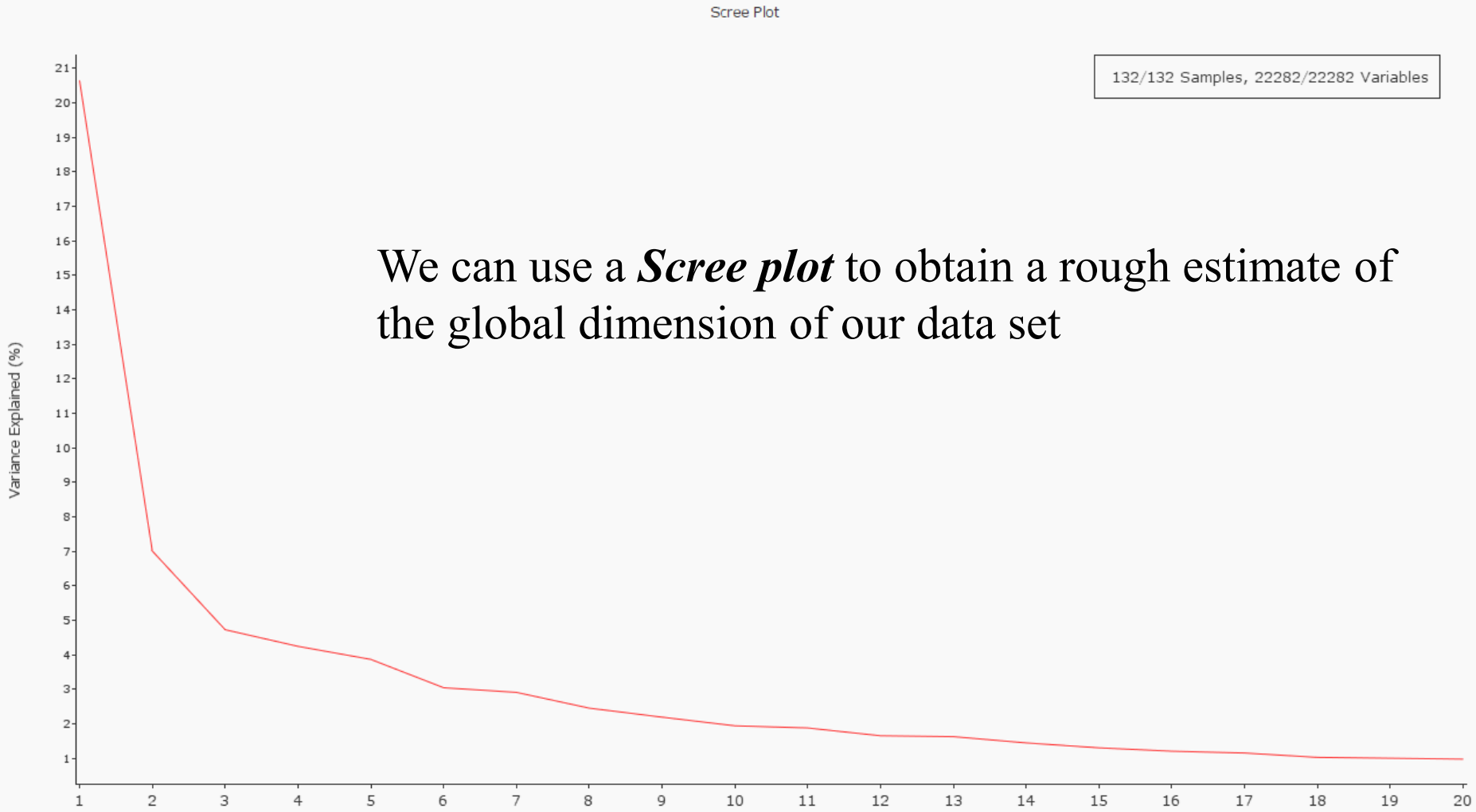


ALL (Leukemia) samples
mRNA bulk

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



We can use a ***Scree plot*** to obtain a rough estimate of the global dimension of our data set

Fontes and Soneson *BMC Bioinformatics* 2011, **12**:307
http://www.biomedcentral.com/1471-2105/12/307

BMC Bioinformatics

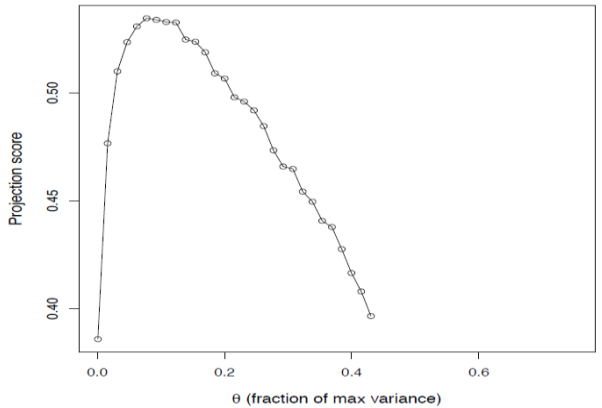
RESEARCH ARTICLE Open Access

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

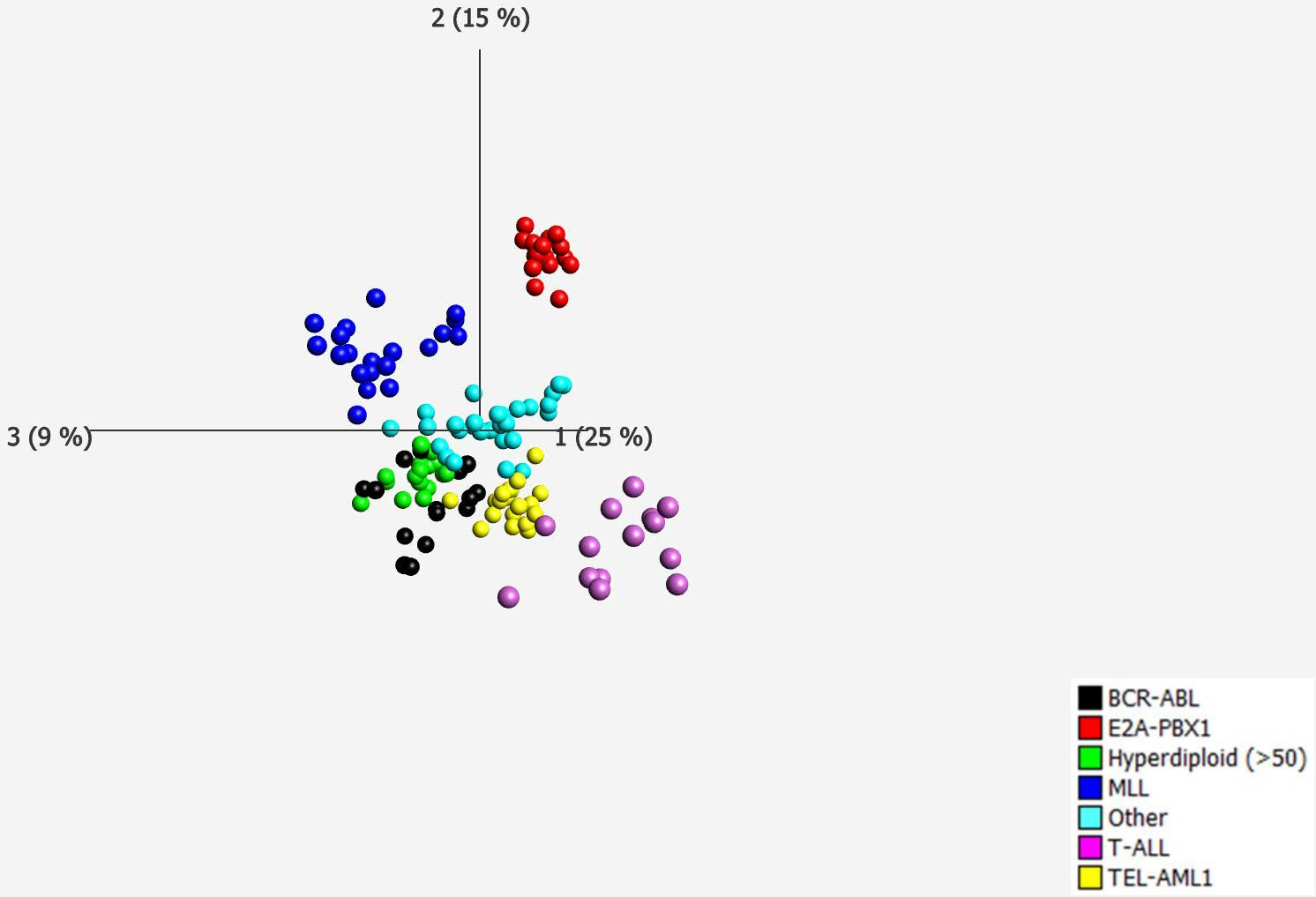
Magnus Fontes* and Charlotte Soneson

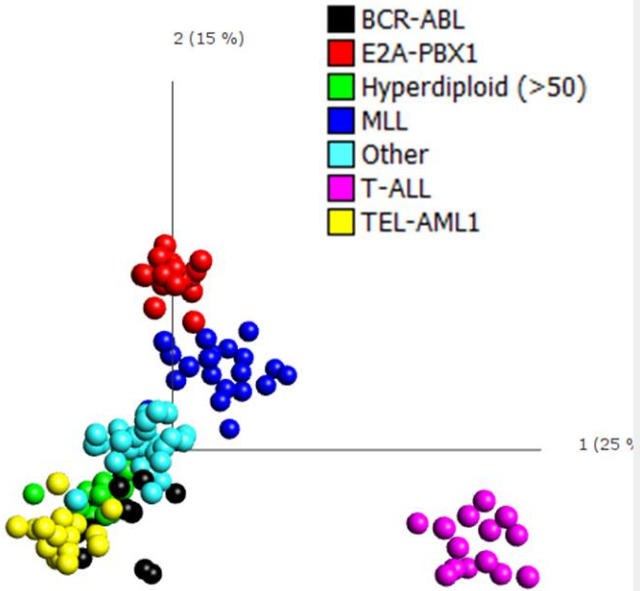
$$\alpha_2(\mathbf{\Lambda}_\mathbf{X}, S) = \frac{\sum_{k\in S}\lambda_k^2}{\sum_{k=1}^{r}\lambda_k^2}.$$

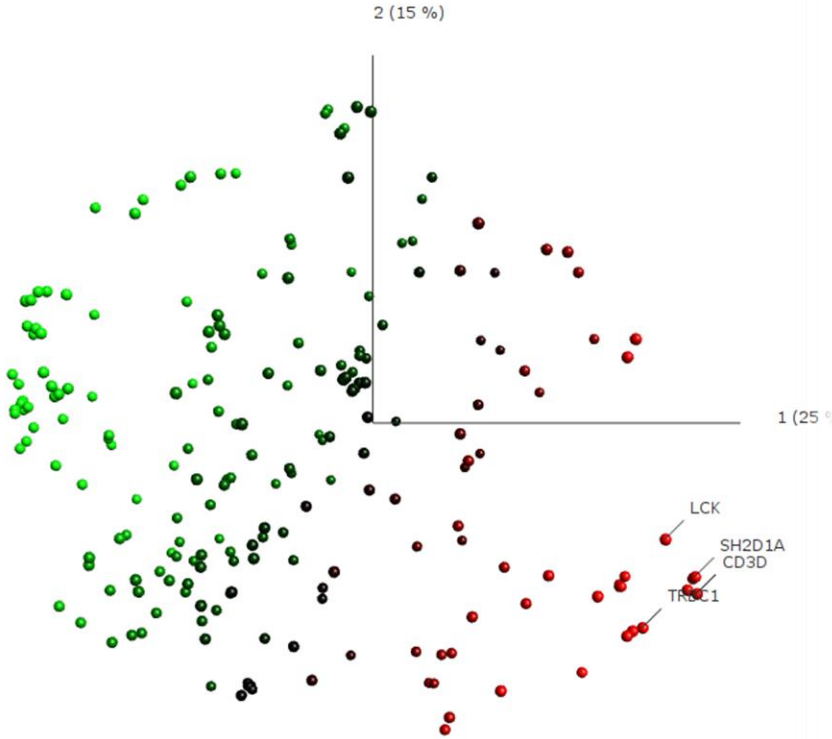
$$\tau(\phi_m(\mathbf{X}), S, \mathcal{P}_{\phi_m(\mathbf{X})}) = \left(\alpha_2(\mathbf{\Lambda}_{\phi_m(\mathbf{X})}, S)\right)^{1/2} - \mathbb{E}_{\mathcal{P}_{\phi_m(\mathbf{X})}}\left[\left(\alpha_2(\mathbf{\Lambda}_{\phi_m(\mathbf{X})}, S)\right)^{1/2}\right].$$



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.


2 (15 %)
3 (9 %)
1 (25 %)
BCR-ABL
E2A-PBX1
Hyperdiploid (>50)
MLL
Other
T-ALL
TEL-AML1



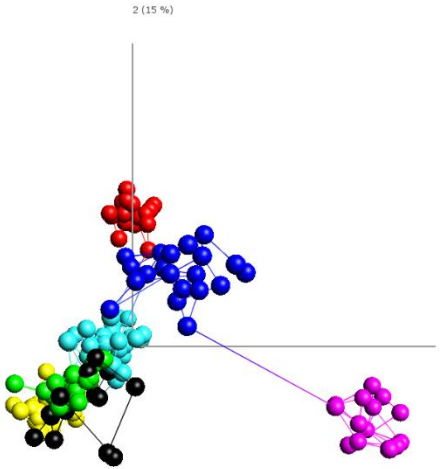


Duality of PCA → *Bi-plot* after optimization over *projection score*

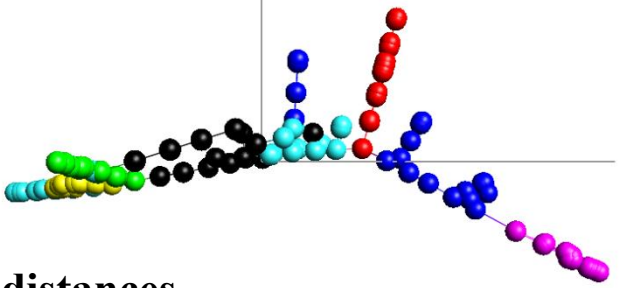
www.sciencemag.org SCIENCE VOL 290 22 DECEMBER 2000

A Global Geometric Framework for Nonlinear Dimensionality Reduction

Joshua B. Tenenbaum,[1]* Vin de Silva,[2] John C. Langford[3]



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



MDS based on graph distances

ISOMAP

RDocumentation

isomap

From vegan v2.4-2 by Jari Oksanen

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=10, ...)
isomapdist(dist, epsilon, k, path = "shortest", fragmentedOK =FALSE, ...)
"summary"(object, axes = 4, ...)
"plot"(x, met = TRUE, n.col = "gray", type = "points", ...)
```

Arguments

dist	Dissimilarities.
ndim	Number of axes in metric scaling (argument `k` in `cmdscale`).
epsilon	Shortest dissimilarity retained.

Stochastic Neighbor Embedding (SNE)

Original distances

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

Original SNE

Sam Roweis and Geoffrey Hinton

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

Distances in reduced space

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

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

Stochastic Neighbor Embedding (SNE) and tSNE

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)

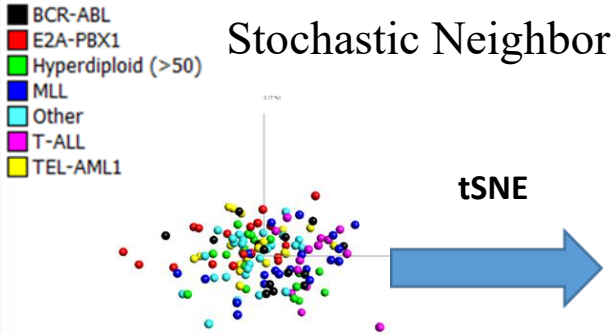
LinkingTo Rcpp

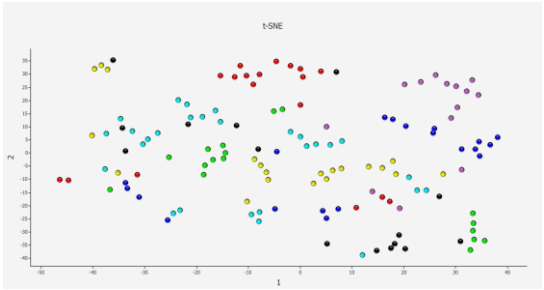
Suggests testthat

RoxygenNote 5.0.1

NeedsCompilation yes

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



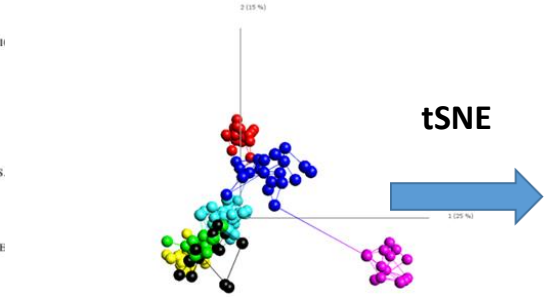


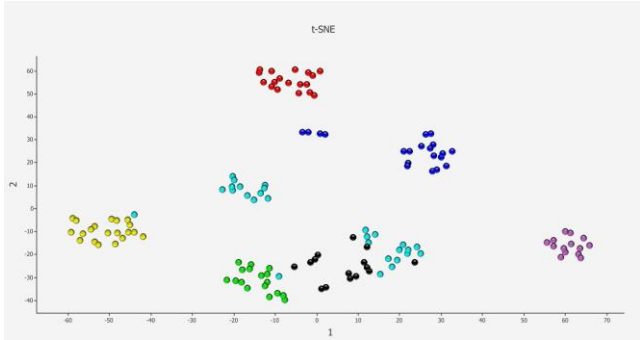
Journal of Machine Learning Research 1 (2008) 1-48 Submitted 4/00; Published 1

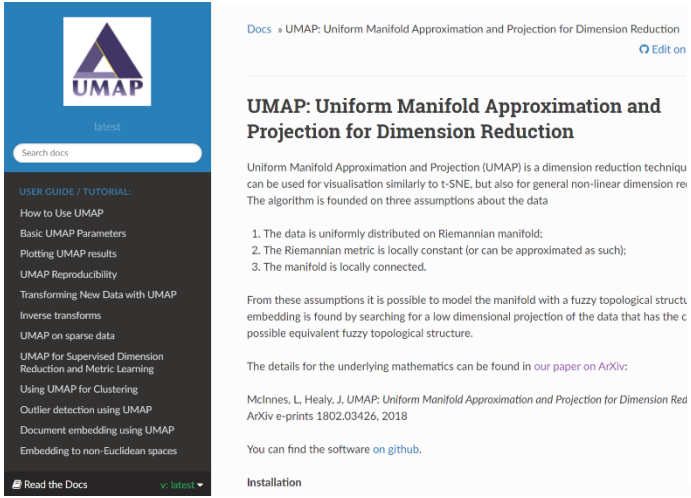
Visualizing Data using t-SNE

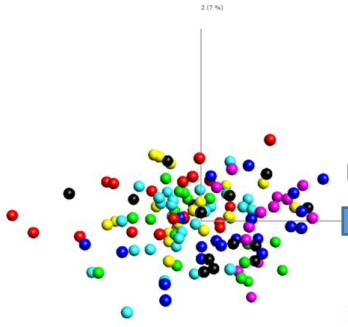
Laurens van der Maaten L.VANDERMAATEN@MICC.UNIMAAS.
MICC-IKAT
Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands

Geoffrey Hinton HINTON@CS.TORONTO.E
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

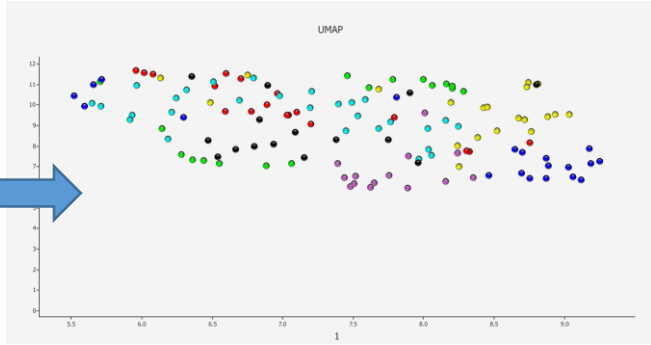


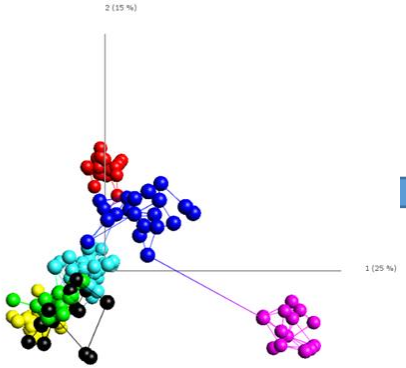



UMAP
latest
Search docs
USER GUIDE / TUTORIAL:
How to Use UMAP
Basic UMAP Parameters
Plotting UMAP results
UMAP Reproducibility
Transforming New Data with UMAP
Inverse transforms
UMAP on sparse data
UMAP for Supervised Dimension Reduction and Metric Learning
Using UMAP for Clustering
Outlier detection using UMAP
Document embedding using UMAP
Embedding to non-Euclidean spaces
Read the Docs
v: latest
Docs » UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
Edit on
UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction techniqu
can be used for visualisation similarly to t-SNE, but also for general non-linear dimension re
The algorithm is founded on three assumptions about the data
1. The data is uniformly distributed on Riemannian manifold;
2. The Riemannian metric is locally constant (or can be approximated as such);
3. The manifold is locally connected.
From these assumptions it is possible to model the manifold with a fuzzy topological structu
embedding is found by searching for a low dimensional projection of the data that has the c
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 Red
ArXiv e-prints 1802.03426, 2018
You can find the software on github.
Installation

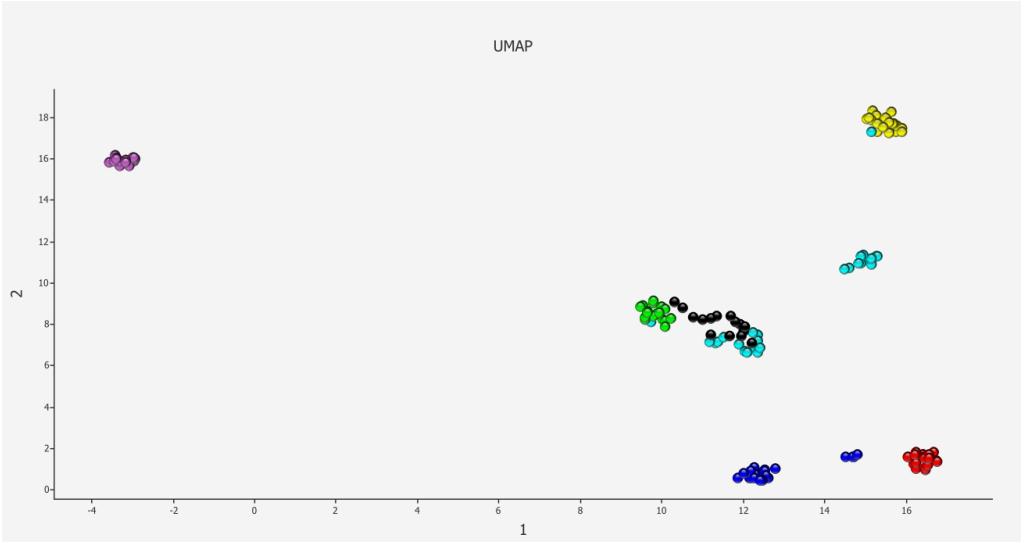


UMAPS

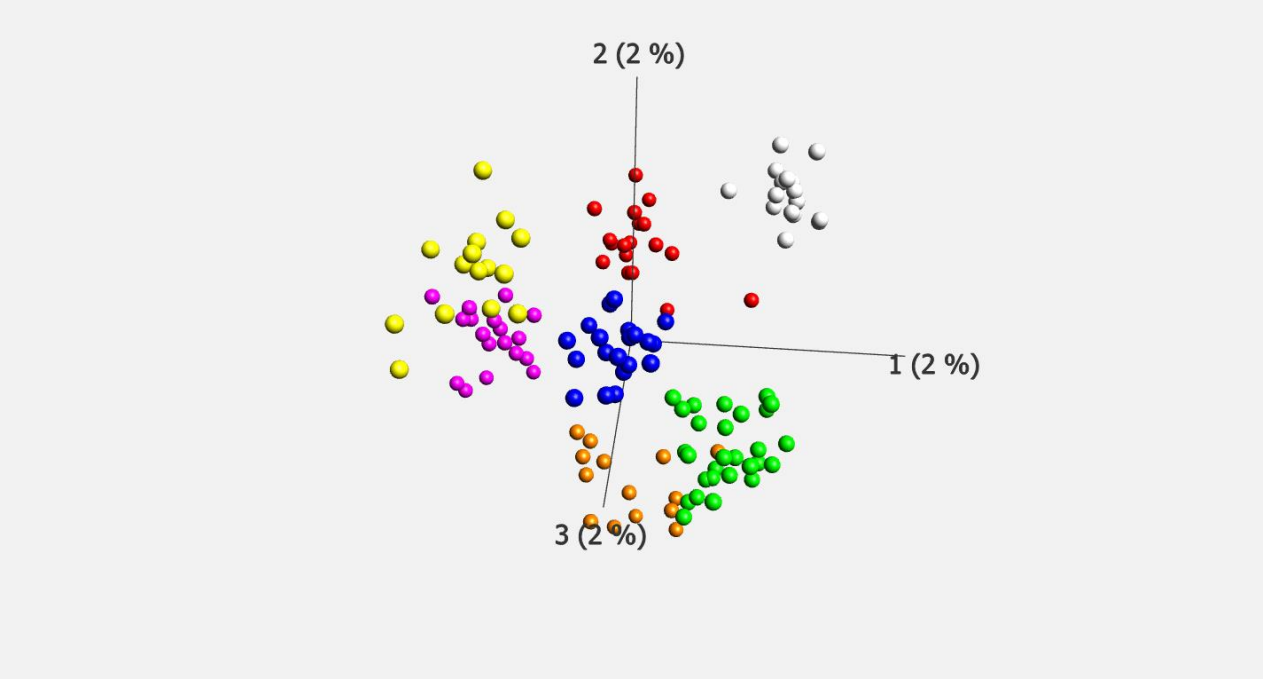

UMAP
1


2 (15 %)
1 (25 %)

UMAPS

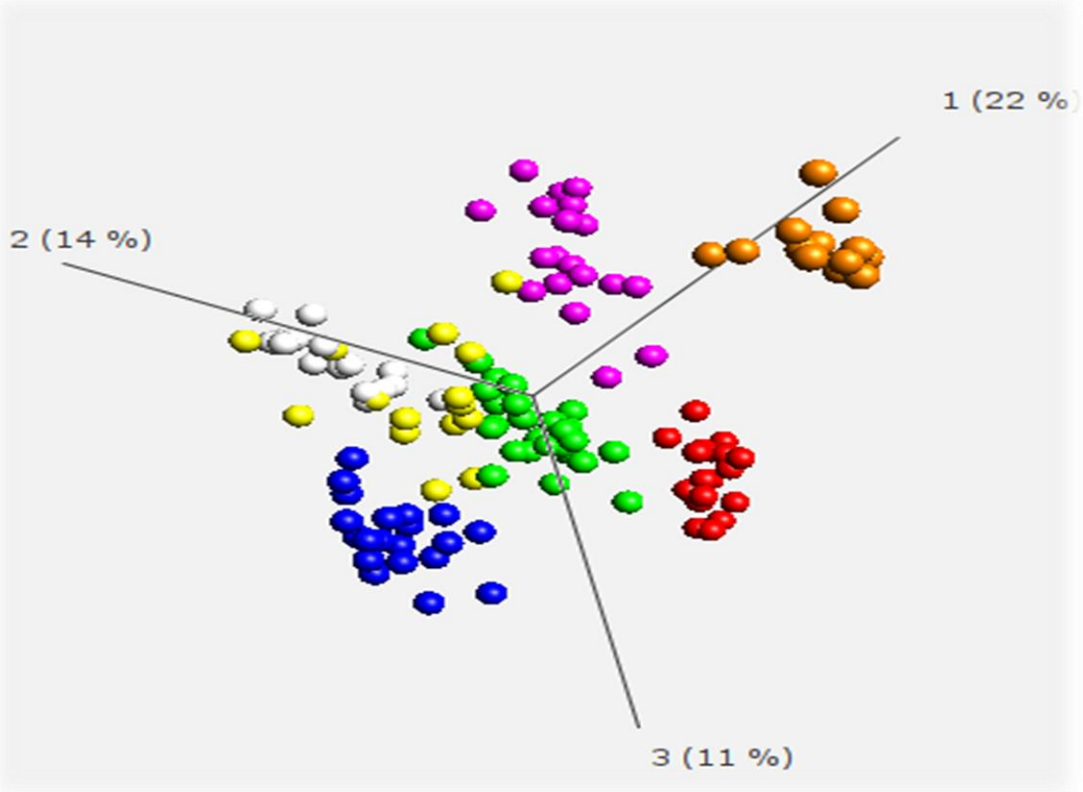

UMAP
2
1

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*22282 =1114.1*

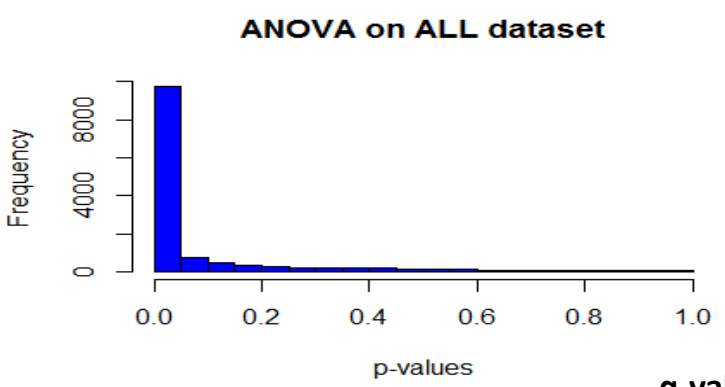
ANOVA on ALL dataset

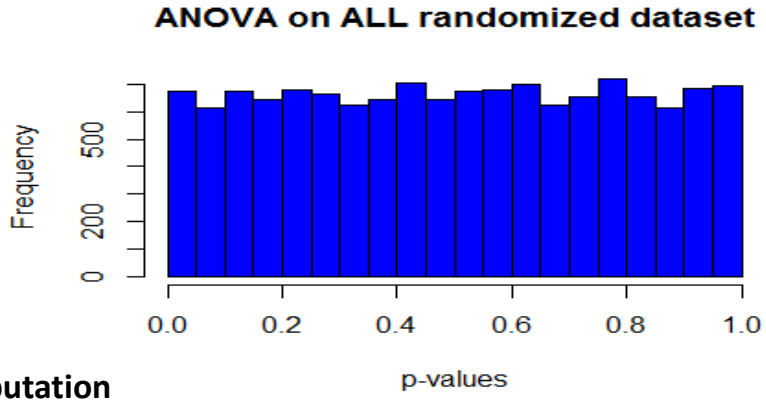


- BCR-ABL
- E2A-PBX1
- Hyperdiploid (>50)
- MLL
- Other
- T-ALL
- TEL-AML1

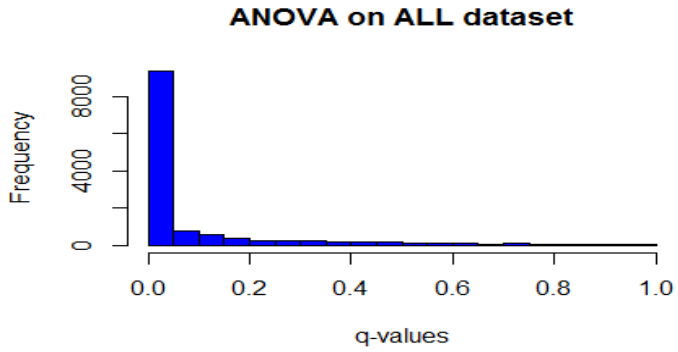
Filtering down to 1000 Variables (p=6e-13)

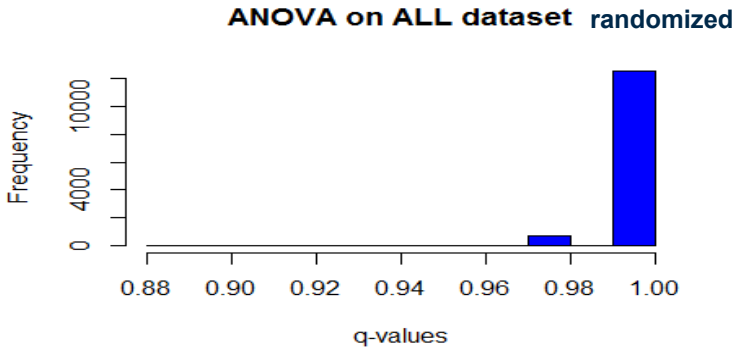
The p-value distributions





q-value computation





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 348 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

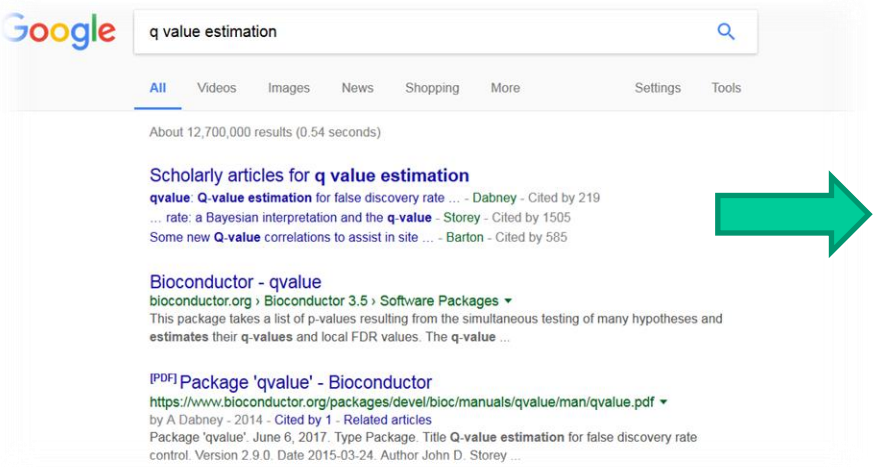
Sure! Here is an example of how to perform principal component analysis (PCA) in R:

```
# 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")
```

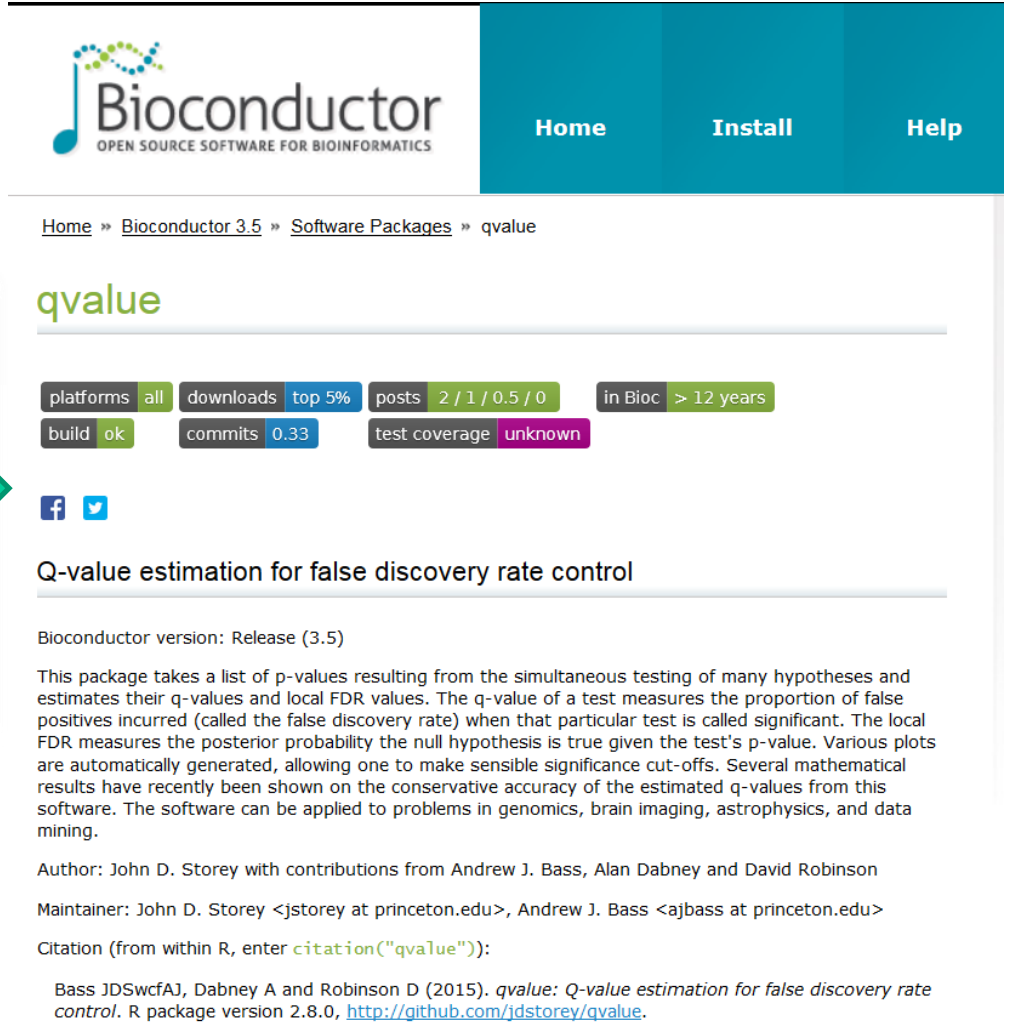
In this example, `data` is a dataframe containing the data we want to perform PCA on. We use the `prcomp()` function to perform the PCA, with the `scale = TRUE` argument indicating that we want to scale the data before performing the analysis. We then use `summary()` to print a summary of the results,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

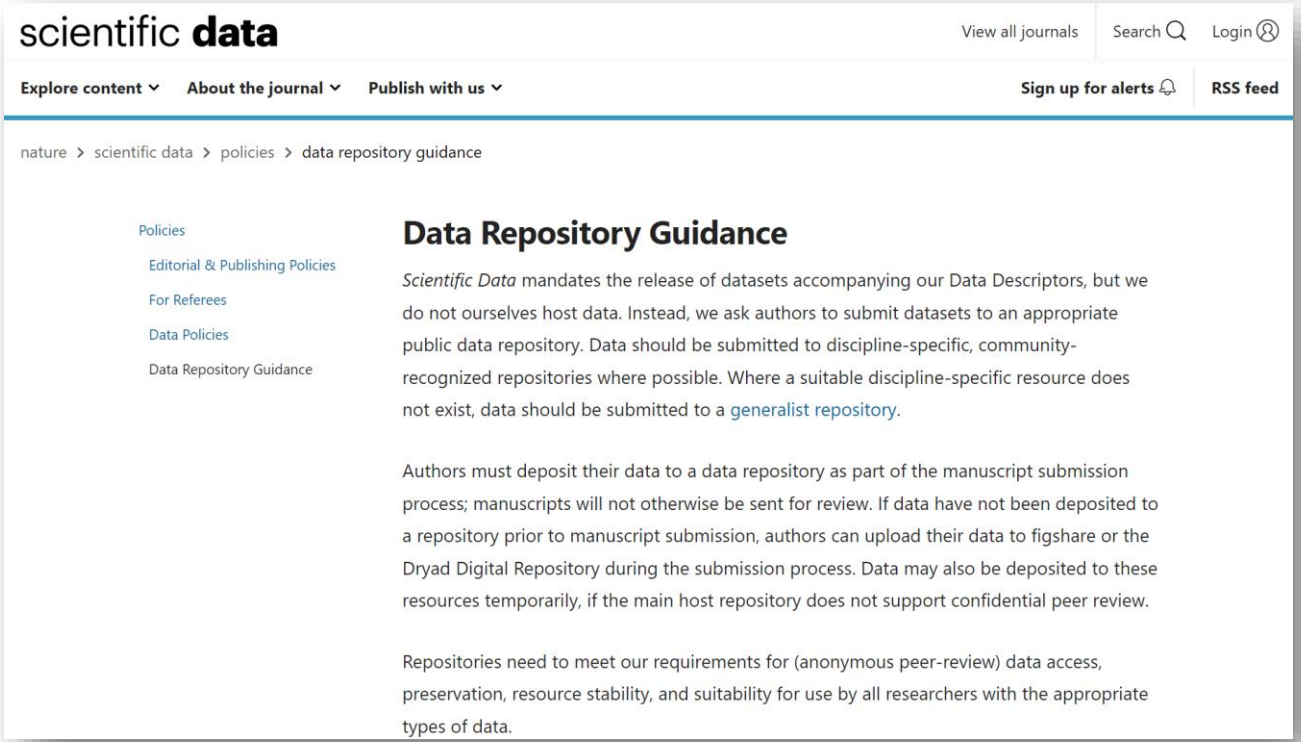
Google search for *q value estimation*



lead to R implementation in *Bioconductor* →

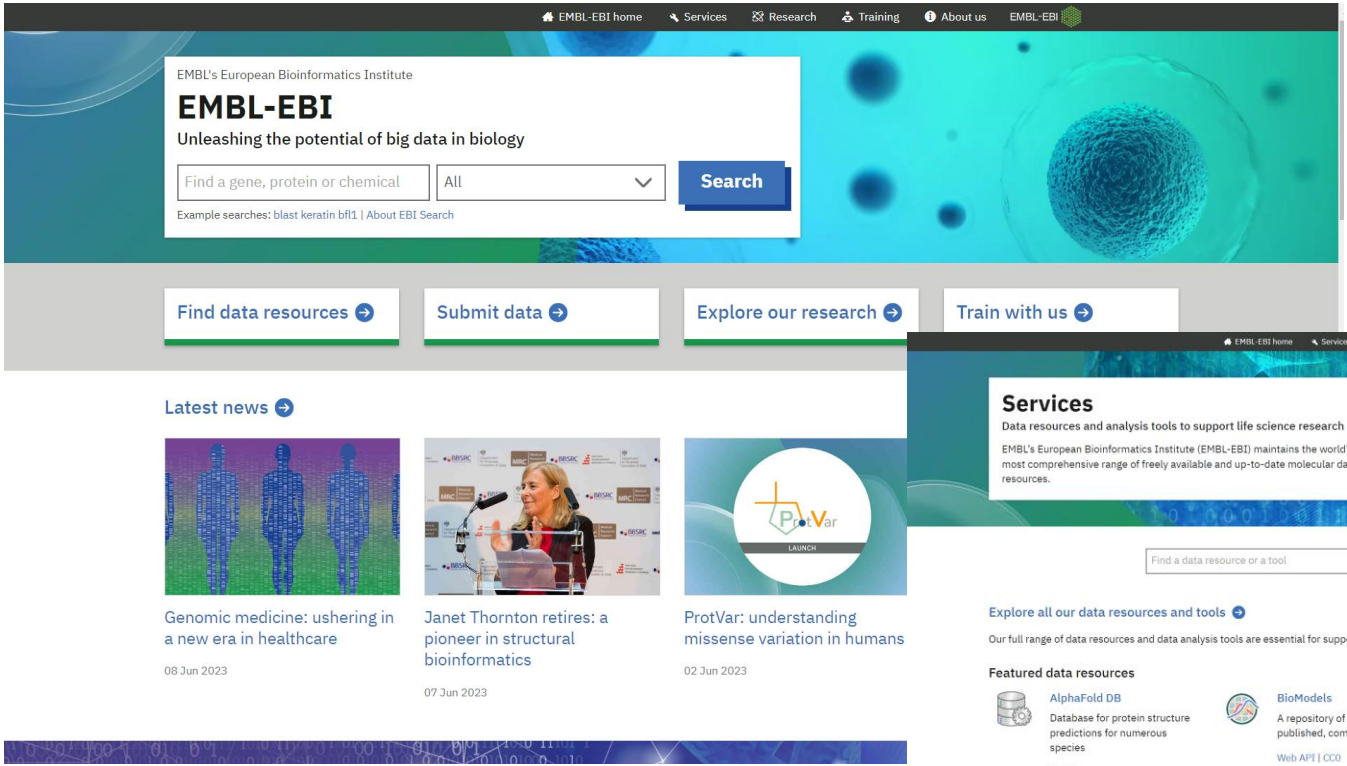


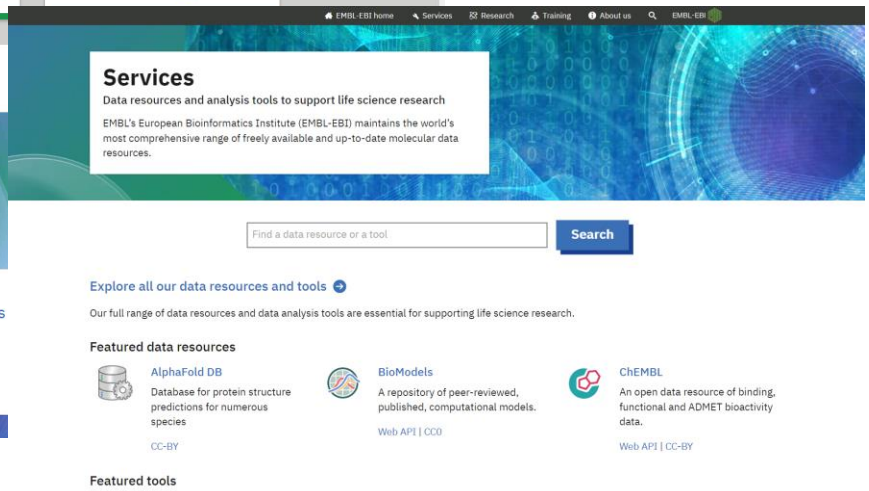
Where to find and deposit biomedical data?



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

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





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

Expression Atlas

Gene expression across species and biological conditions

Query single cell expression

To Single Cell Expression Atlas

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

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

Search | Gene set enrichment

Gene / Gene properties

Enter gene query…

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

Species

Any

Biological conditions

Enter condition query...

Examples: lung, leaf, valproic acid, cancer

Search | Clear

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

ega-archive.org

EUROPEAN GENOME-PHENOME ARCHIVE

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...

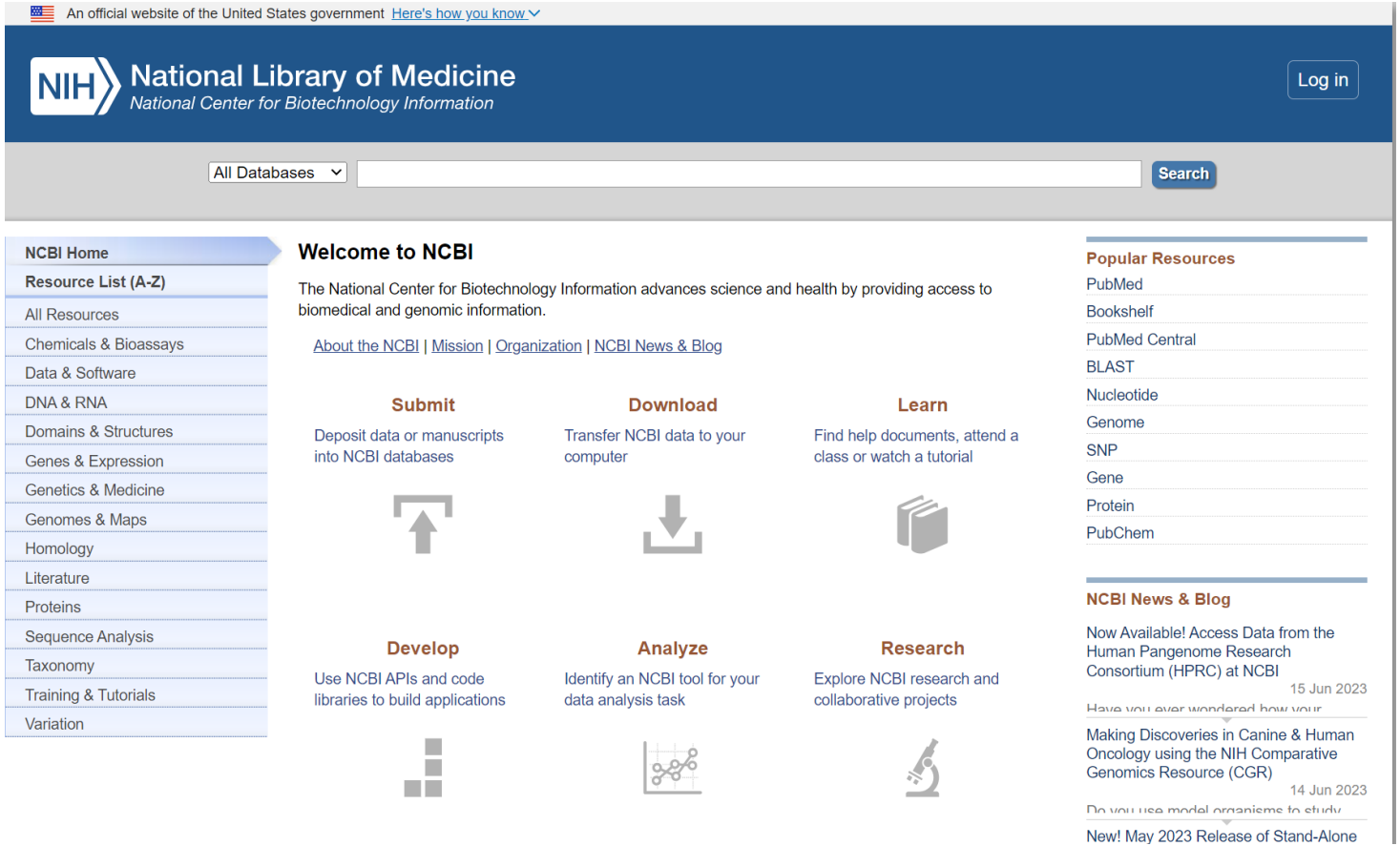
DOWNLOAD DATA

DEPOSIT DATA

MANAGE DATA

https://ega-archive.org/

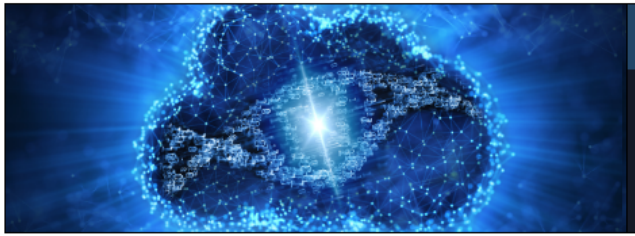
https://www.ncbi.nlm.nih.gov/



SRA

SRA | Search

Advanced | Help



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

You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA)

Support Center

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

NCBI

DATASET CURATED BROWSER

Gene Expression Omnibus

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 20

4348 DataSet records

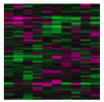
Page 1 of 218 > >>

DataSet	Title	Organism(s)	Platform	Series	Samples
GDS6063	Influenza A effect on plasmacytoid dendritic cells	*Homo sapiens*	GPL10558	GSE68849	10
GDS6010	Influenza virus H5N1 infection of U251 astrocyte cell line: time course	*Homo sapiens*	GPL6480	GSE66597	18
GDS5879	Pulmonary CDC11c+ cells from young and middle-age animals	*Mus musculus*	GPL6885	GSE71868	8
GDS5826	Multiple myeloma cell lines with acquired resistance to chemotherapeutic agent carfilzomib	*Homo sapiens*	GPL570	GSE69078	12
GDS5825	Interleukin-1α deficiency effect on injured spinal cord	*Mus musculus*	GPL6246	GSE70302	12
GDS5881	Nebulin deficiency effect on the soleus	*Mus musculus*	GPL6246	GSE70213	12
GDS5880	Nebulin deficiency effect on the quadriceps	*Mus musculus*	GPL6246	GSE70213	12
GDS5913	SRPIN803 small molecule inhibitor of SRPK1 effect on retinal pigment epithelial cell line	*Homo sapiens*	GPL570	GSE62947	6
GDS5665	Pathogen-associated molecular-pattern curdlan effect on interleukin-2 deficient GM-CSF myeloid dendritic cells	*Mus musculus*	GPL6246	GSE58120	12
GDS5662	Histone demethylase KDM3A-deficiency effect on estrogen-stimulated breast cancer cells in vitro	*Homo sapiens*	GPL10558	GSE68918	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:	*Homo sapiens*		
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. *J Immunol* 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

NLM NIH GEO Help Disclaimer Accessibility

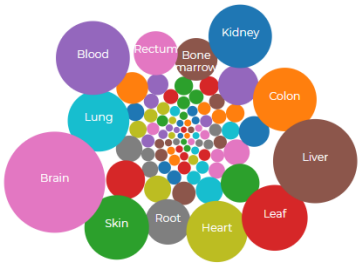
https://www.ncbi.nlm.nih.gov/geo/

Organism, repository, gene, tissue, accession

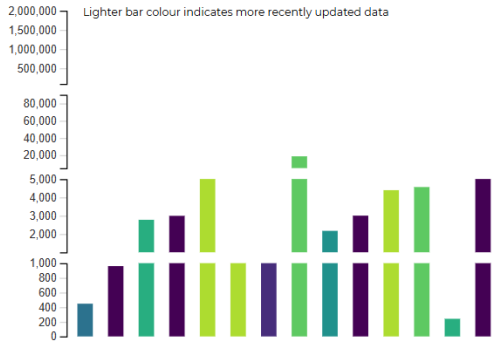
Advanced Search

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

Description Sample Data



Tissues Organisms Diseases



Resources Omics

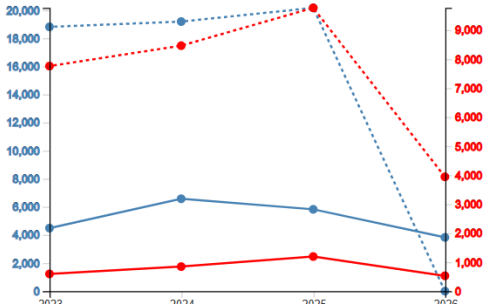
Latest Datasets

- May 31 '26 Effects of different light conditions in Centella a...
- May 31 '26 4â€²-O-(L-Alanylated) DMC as a EEbB3 Inhibito...
- May 30 '26 Implant renal injury-responsive cells to supple...
- May 30 '26 Multiomic profiling of vascular endothelial cell ...
- May 30 '26 SoFIC Orchestrates Physical Barriers and Meta...
- May 30 '26 Mouse MafA Cut&Run in Min6 cells
- May 30 '26 CMPortal: a community-oriented, data-driven r...
- May 30 '26 Sorbate induces lysine sorbylation through no...
- May 30 '26 Virus-inclusive single-nucleus RNA sequencing...
- May 30 '26 Single-nucleus RNA sequencing elucidates the...

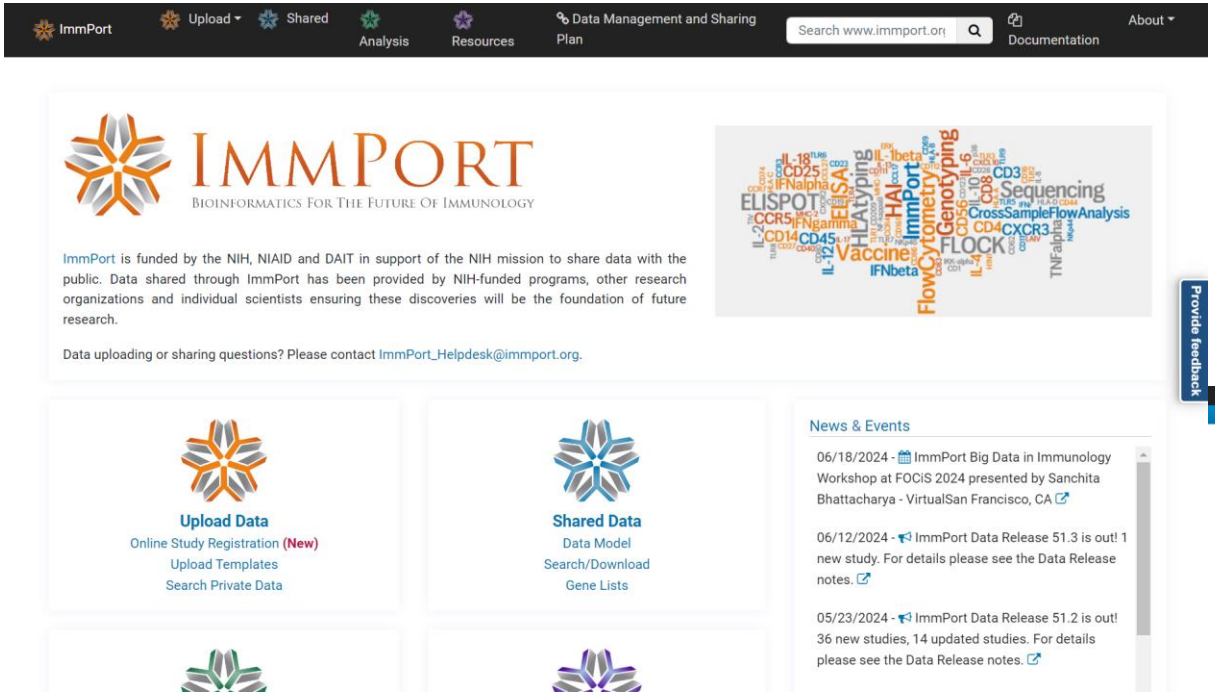
Most Accessed Datasets

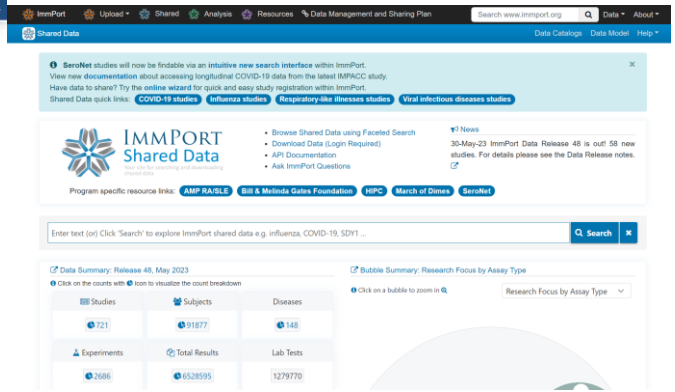
- 15787 Kholodenko1999 - EGFR signaling
- 12772 Ndairou2020 - early-stage transmission dynamics ...
- 9440 A metabolomic study of urinary changes in type 2 ...

New Datasets Per Year



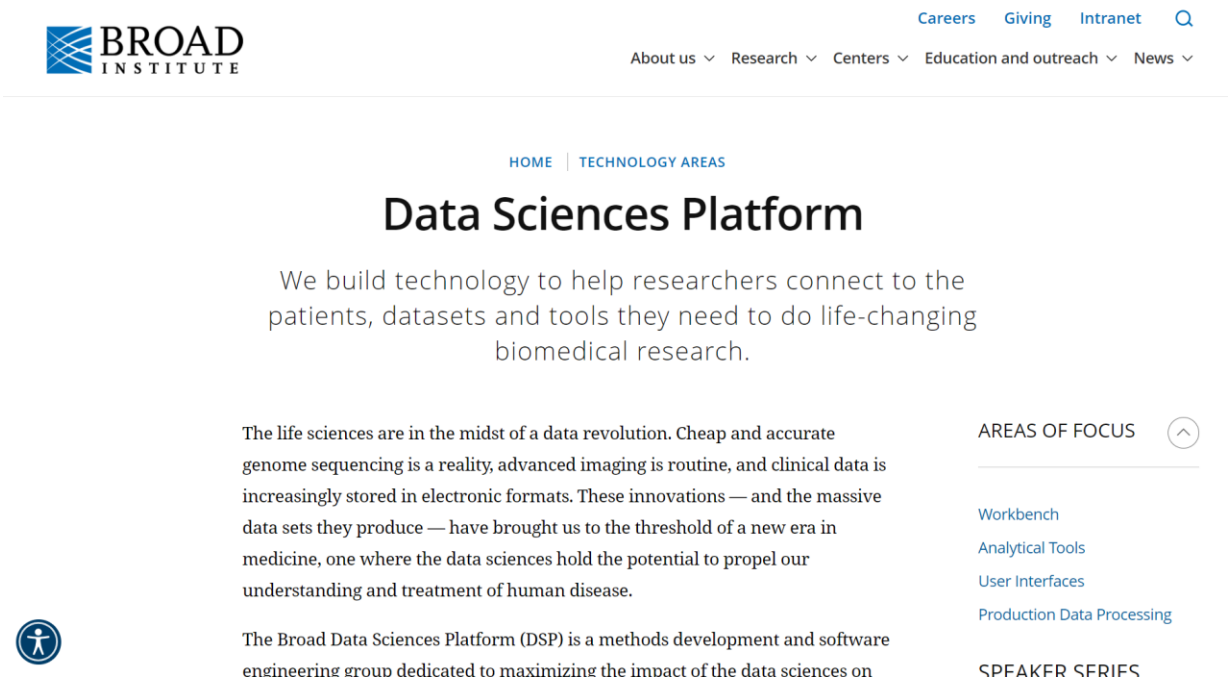
https://www.immport.org/home





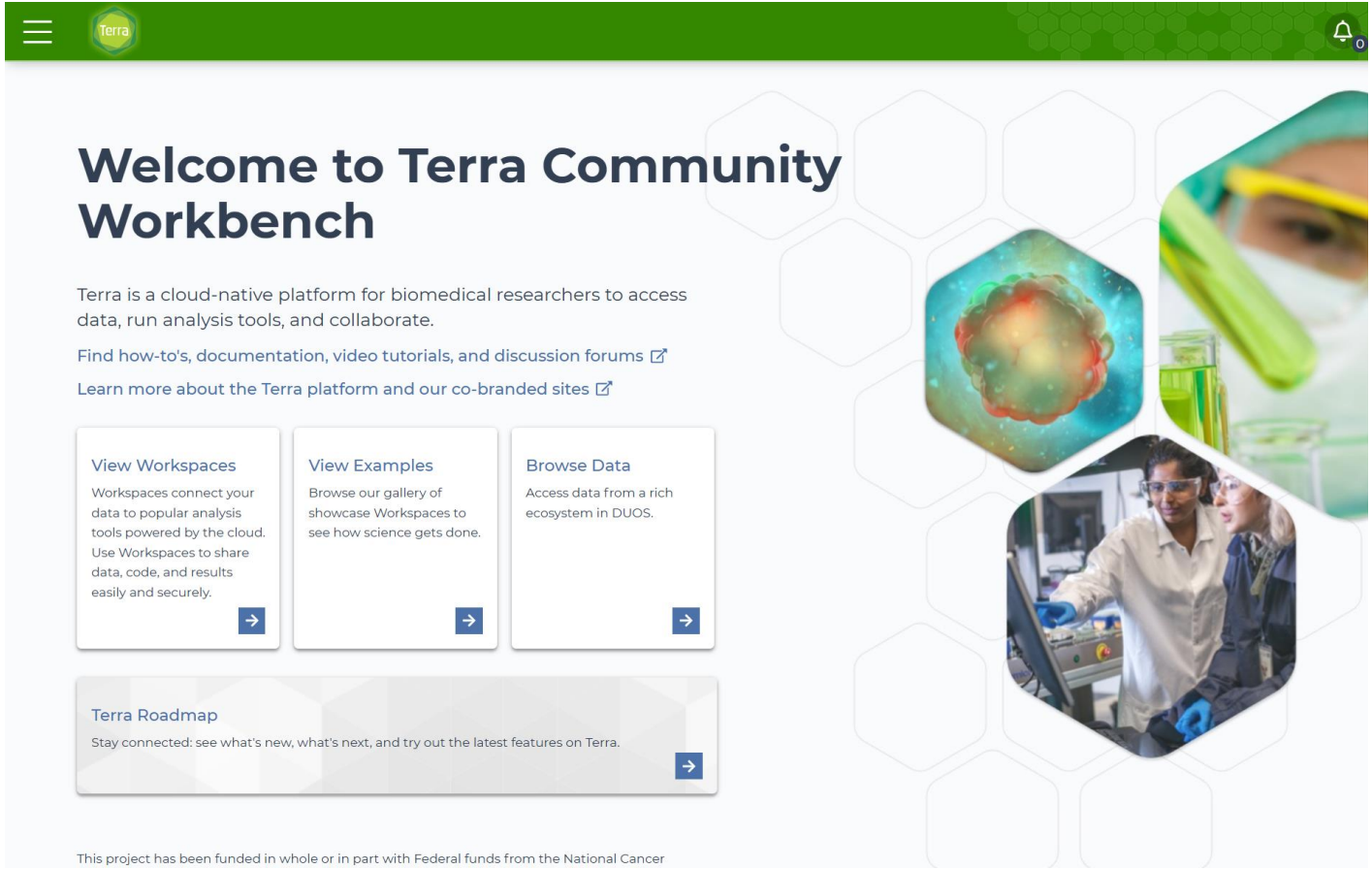
https://www.immport.org/shared/home

https://www.broadinstitute.org/

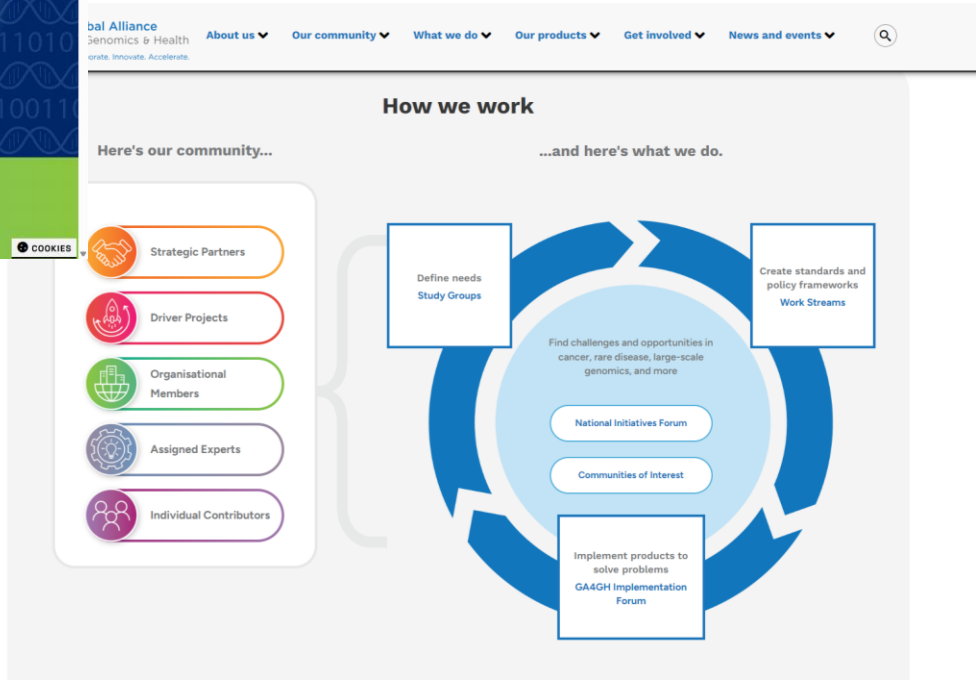


https://www.broadinstitute.org/data-sciences-platform

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



https://www.ga4gh.org/



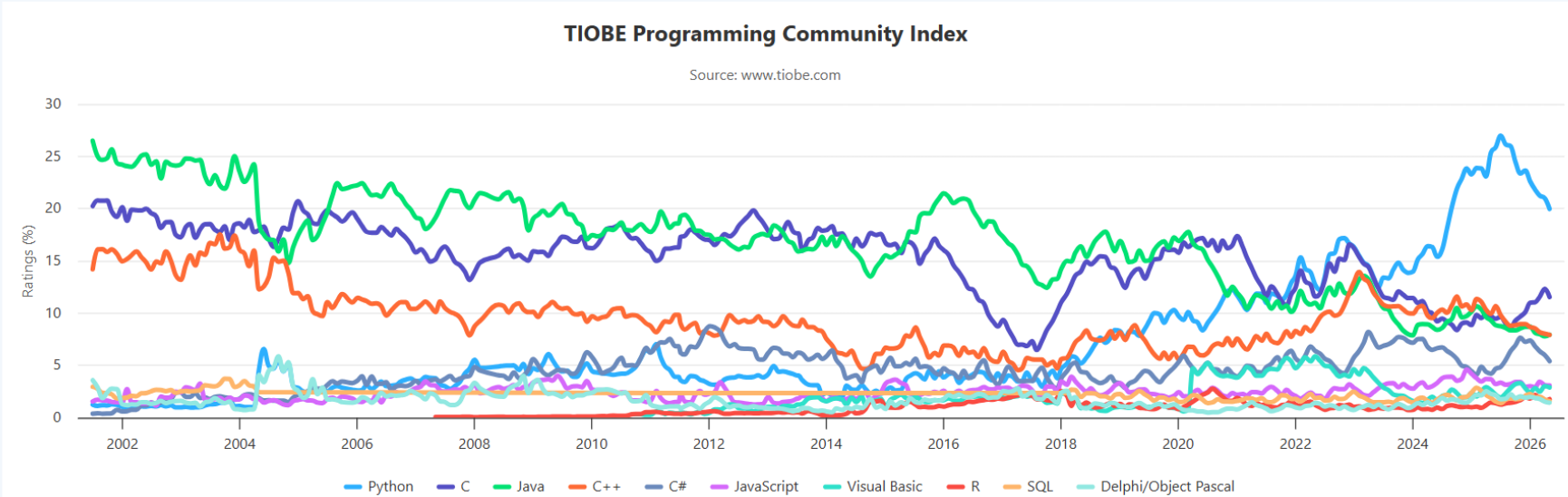
Formed in 2013, the Global Alliance for Genomics and Health (GA4GH) unites an international community dedicated to advancing human health through genomic data. We build technical standards and policy frameworks and tools that will expand responsible, voluntary, and secure use of genomic and other related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

Interacting with Data - Software tools

TIOBE
the software quality company

About us | Knowledge | News | Coding Standards | TIOBE Index | Contact

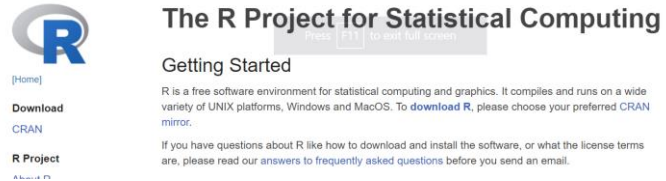
Products | Quality Models | Markets | Schedule a demo



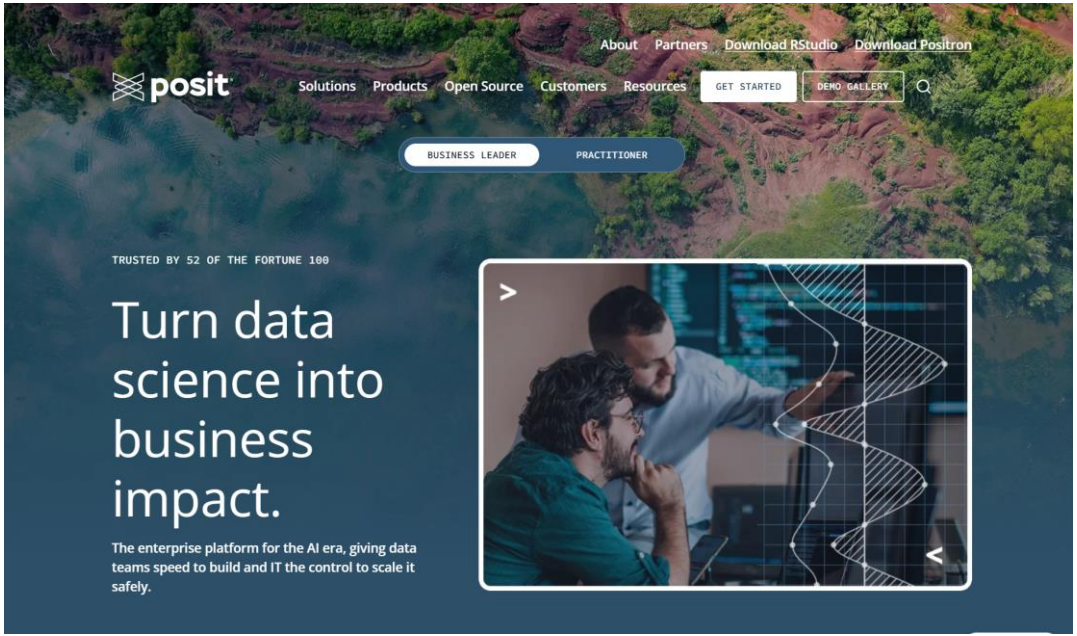
https://www.tiobe.com/tiobe-index/

What tools do I use?

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



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



https://www.bioconductor.org/

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.

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

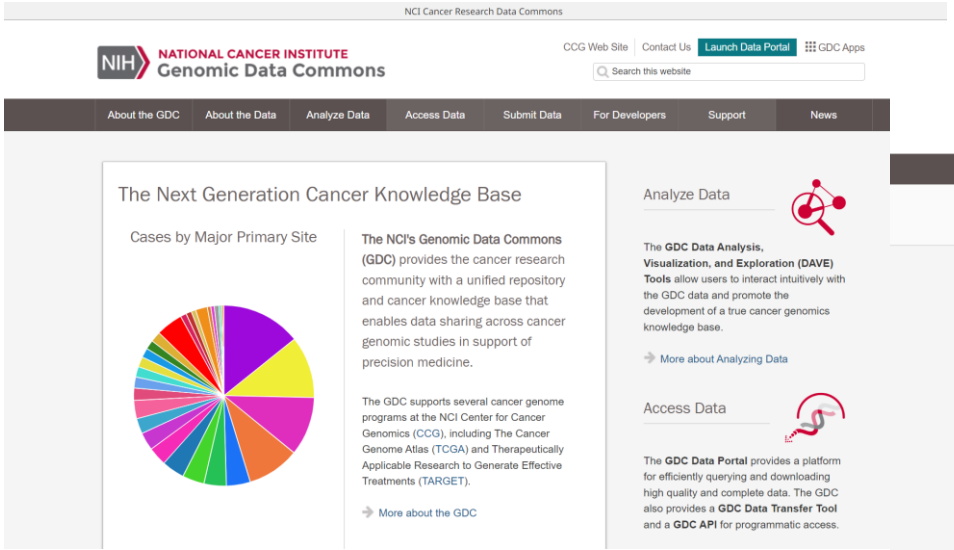
Develop »

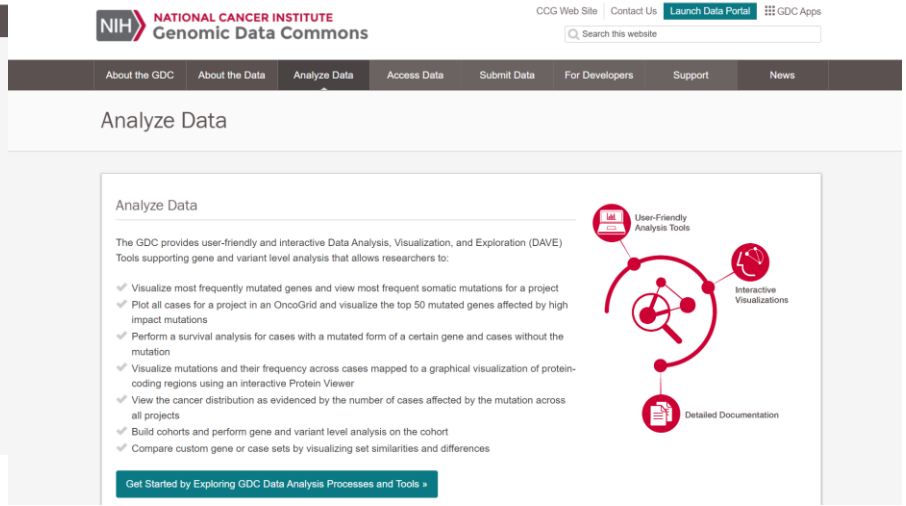
Contribute to *Bioconductor*

- Developer resources
- Use Bioc 'devel'

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

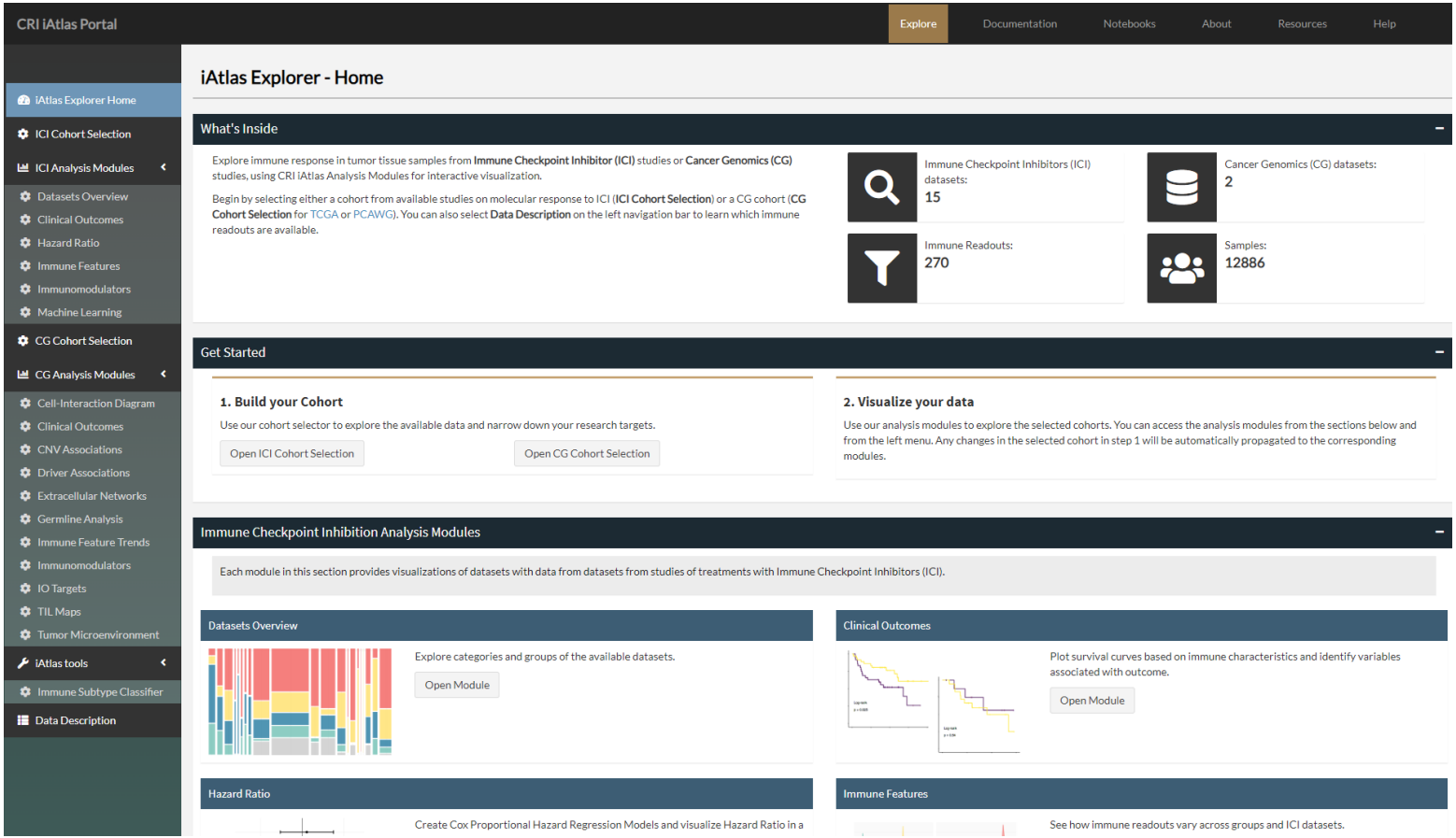
https://gdc.cancer.gov/





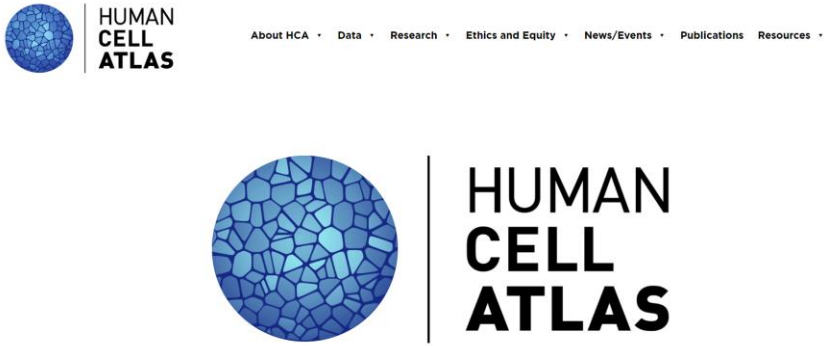
https://gdc.cancer.gov/analyze-data

https://cri-iatlas.org/

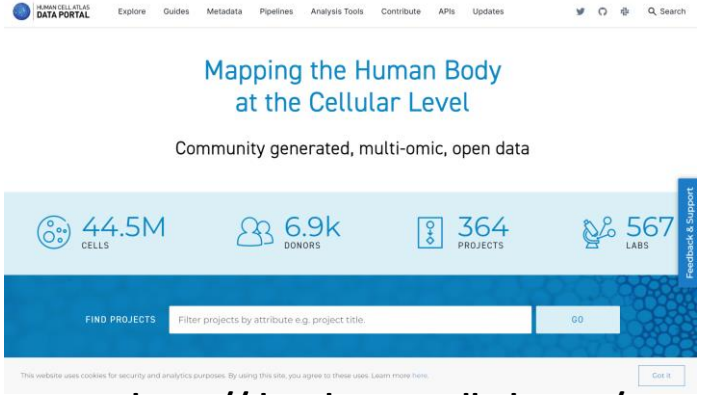


https://isb-cgc.shinyapps.io/iatlas/

Focusing on Cells – The basic units of Life



https://www.humancellatlas.org/



https://data.humancellatlas.org/

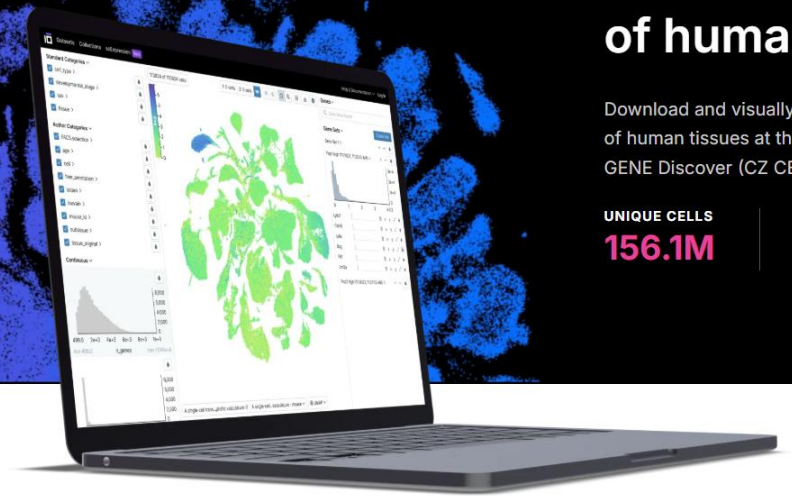
On June 15th 2023 Aviv Regev received the L'Oreal-Unesco for women in science award *"for her pioneering work applying mathematics and computer science to revolutionize cell biology."*

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	DATASETS	CELL TYPES
156.1M	2124	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

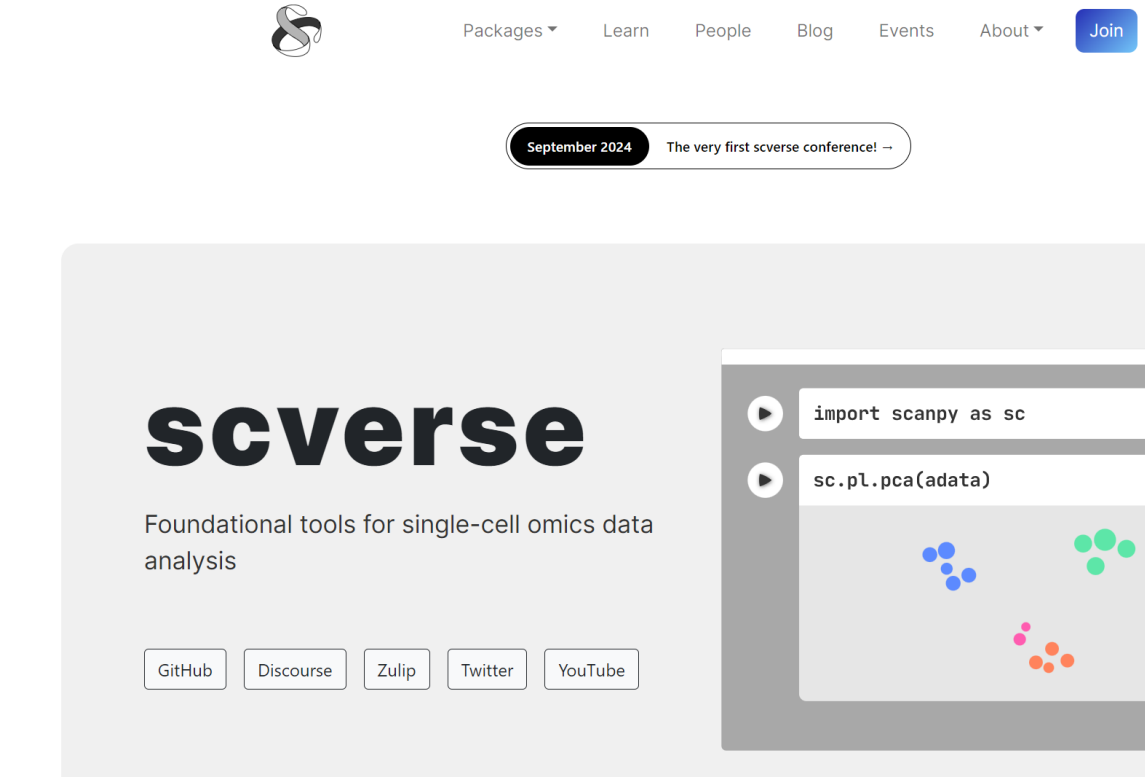
Genotype samples: 837

CyTOF

Cell count: 2,278,550

Flow Cytometry

Cell count: 192,925,633



scverse

A library of single cell analysis algorithms and computational models

https://scverse.org/

MISSION

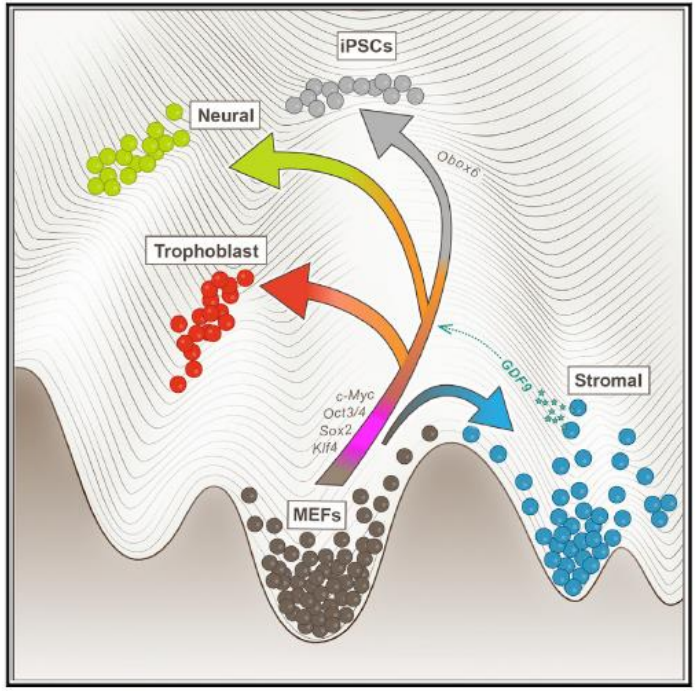
scverse is a consortium of foundational tools (mostly in Python) for omics data in life sciences. It has been founded to ensure the long-term maintenance of these core tools.

Resource

Cell

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

Graphical Abstract



Authors

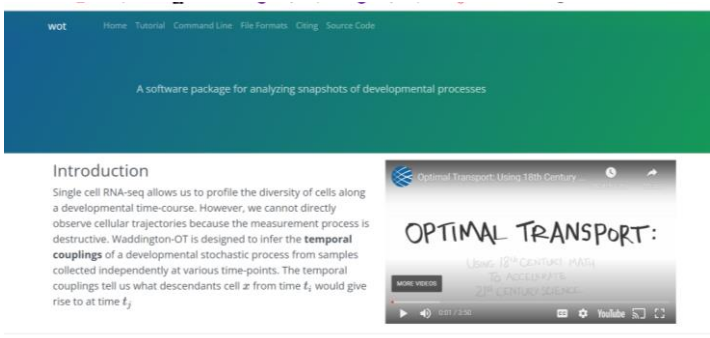
Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, ..., Rudolf Jaenisch, Aviv Regev, Eric S. Lander

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/