AI Safety and the Life Sciences

Gopal P. Sarma MD, PhD

Models, Inference, and Algorithms Initiative

Broad Institute of MIT and Harvard

Complex Interplay Between AI and the Life Sciences

- **Claim 1:** The life sciences will drive the next decade of creativity in AI
  - Economic and social incentives
  - Rich multi-modal data scaling at high exponential rates
  - Challenging biomedical problems

- **Claim 2:** Integrating AI safety and the life sciences is critical and will require significant transdisciplinary efforts

BROAD INSTITUTE

Complex Interplay Between AI and the Life Sciences

- **Claim 1:** The life sciences will drive the next decade of creativity in AI
  - Economic and social incentives

BROAD INSTITUTE

Deloitte.

2019 Global Life Sciences Outlook

Focus and transform | Accelerating change in life sciences

Global health care spending continues to increase dramatically

is projected to reach

$10.059 trillion by 2022


Appian
THE KEY TO SUCCESSFUL
DIGITAL TRANSFORMATION
IN LIFE SCIENCES


Deloitte.
TALES OF
TRANSFORMATION
Regulatory’s impact in life sciences
digital transformation

Digital Transformation for Life Sciences

Cisco Healthcare is empowering innovation. Our technologies make it possible to improve customer engagement, reduce discovery and development time, transform the global supply chain, and address changes in the global and regulatory environment, all while enabling security and compliance.

R&D and Clinical Trials

- Collaboration – video, voice, and virtual meetings
- Cloud and data center
- Data virtualization
- Analytics
- Asset management
- Customer-contact center

Manufacturing and Supply Chain

- Unified, converged factory network
- Connected machines and edge computing
- Centralized, pervasive wireless
- Analytics
- Services-exchange platform
- Asset management

Marketing and Sales Engagement

- Collaboration – video, voice, and virtual meetings
- Automated response bots
- Customer-contact center
- Analytics

Security and DNA for Life Sciences


A GUIDE FOR LIFE SCIENCES COMPANIES
How to Embrace Digital Transformation
READ NOW

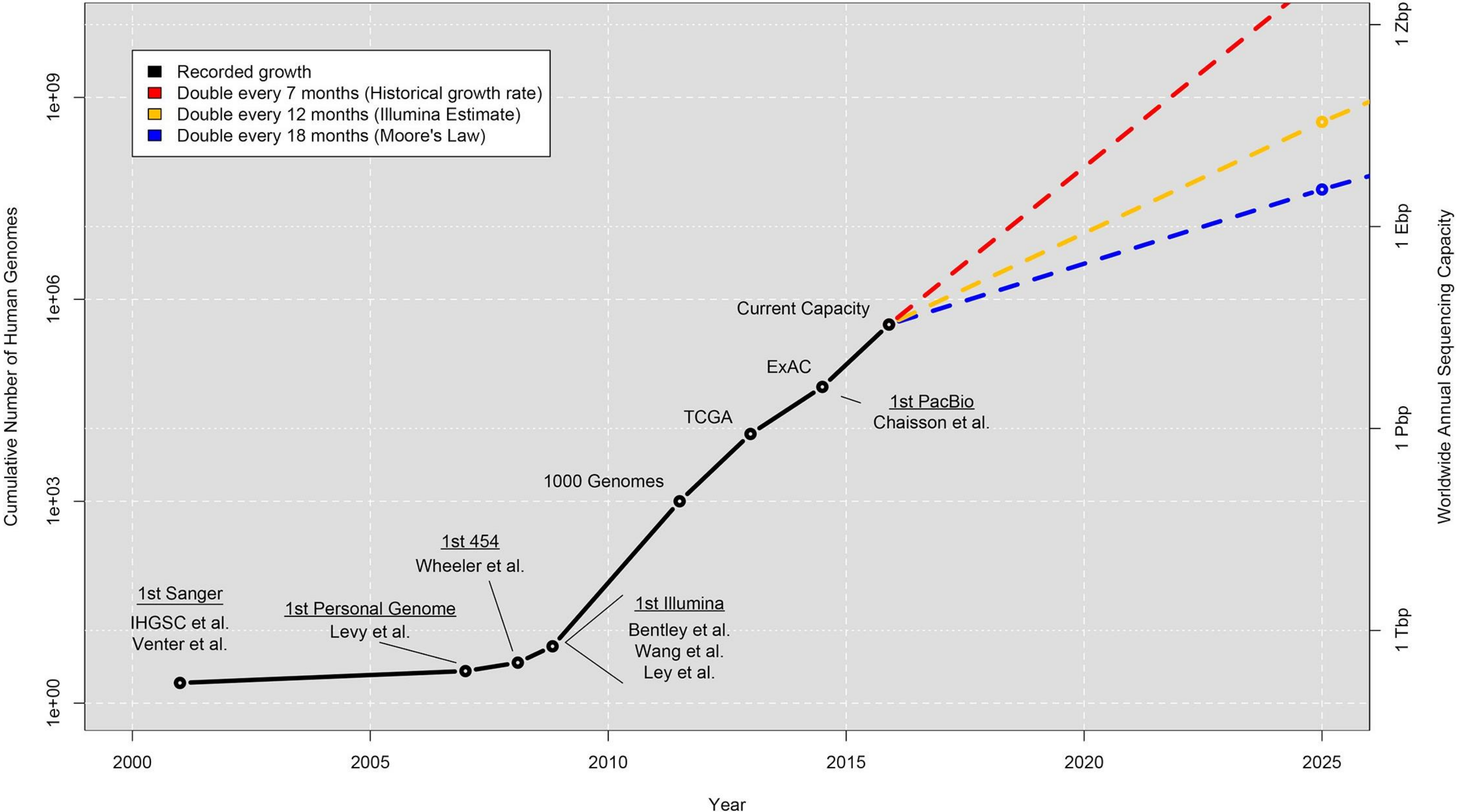

Top 3 Digital Transformation
Challenges in Life Sciences
LEARN MORE
ADLIB

Complex Interplay Between AI and the Life Sciences

- **Claim 1:** The life sciences will drive the next decade of creativity in AI
  - Economic and social incentives
  - Rich multi-modal data scaling at high exponential rates
  - Challenging biomedical problems

Scaling of DNA Sequencing

Stephens ZD et al. (2015) Big Data: Astronomical or Genomical? PLOS Biology 13(7): e1002195.



Bio-inspired AI: Uncovering the Genomic Basis of Disease

hail

https://hail.is

- Open-source, genomic analysis platform
- MatrixTable: first-class support for multi-dimensional structured data, built from the ground-up as a distributed system running in the cloud
- The structure, scale, and analyses of genomic data inspired a new distributed data structure that is broadly useful in biology and beyond

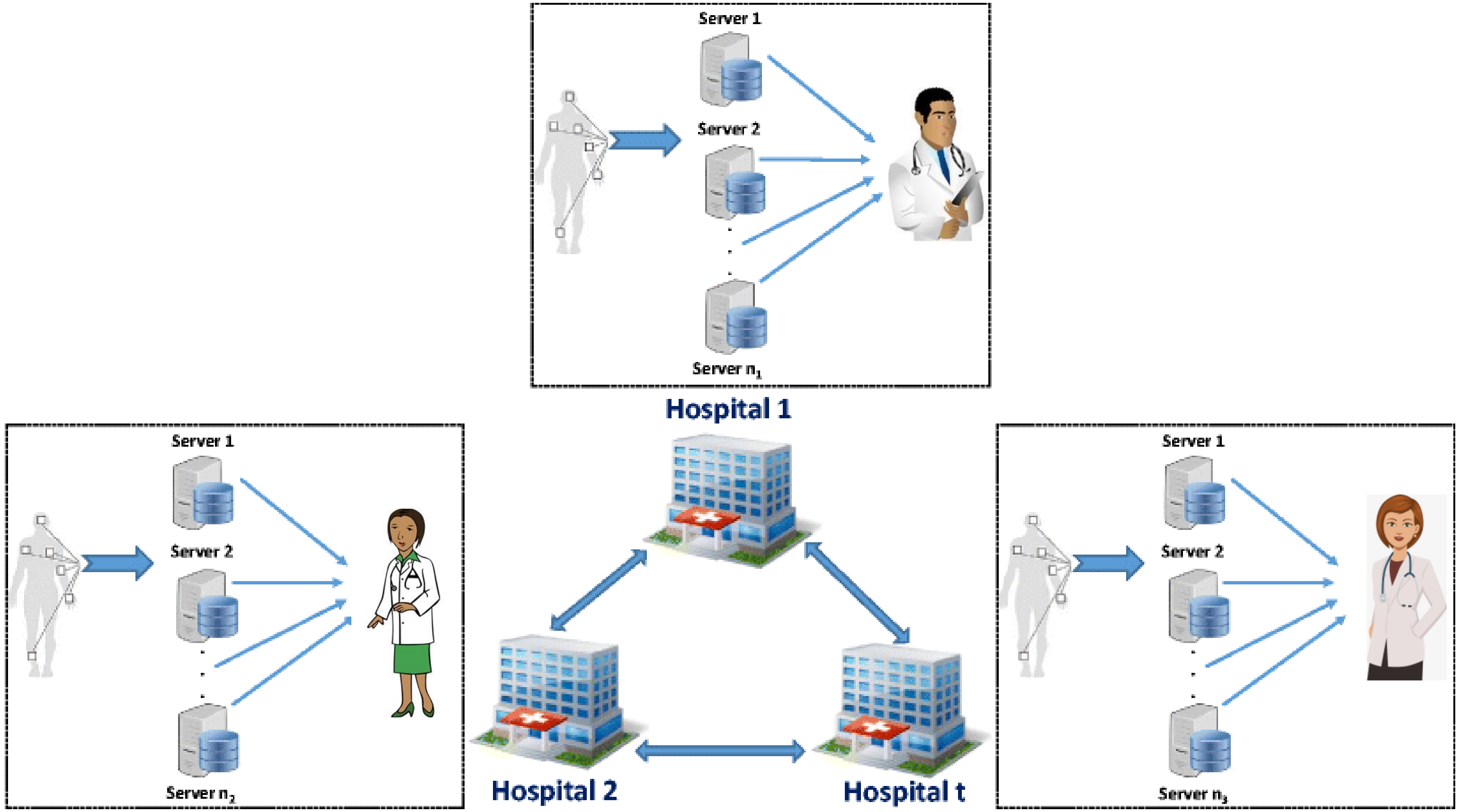
BROAD INSTITUTE

Bio-Inspired AI: Secure Multi-Party Linear Regression

J. Bloom "Secure multi-party linear regression at plaintext speed." *arXiv preprint* arXiv:1901.09531 (2019).
B. Berger and H. Cho "Emerging technologies towards enhancing privacy in genomic data sharing" *Genome Biology* (2019)

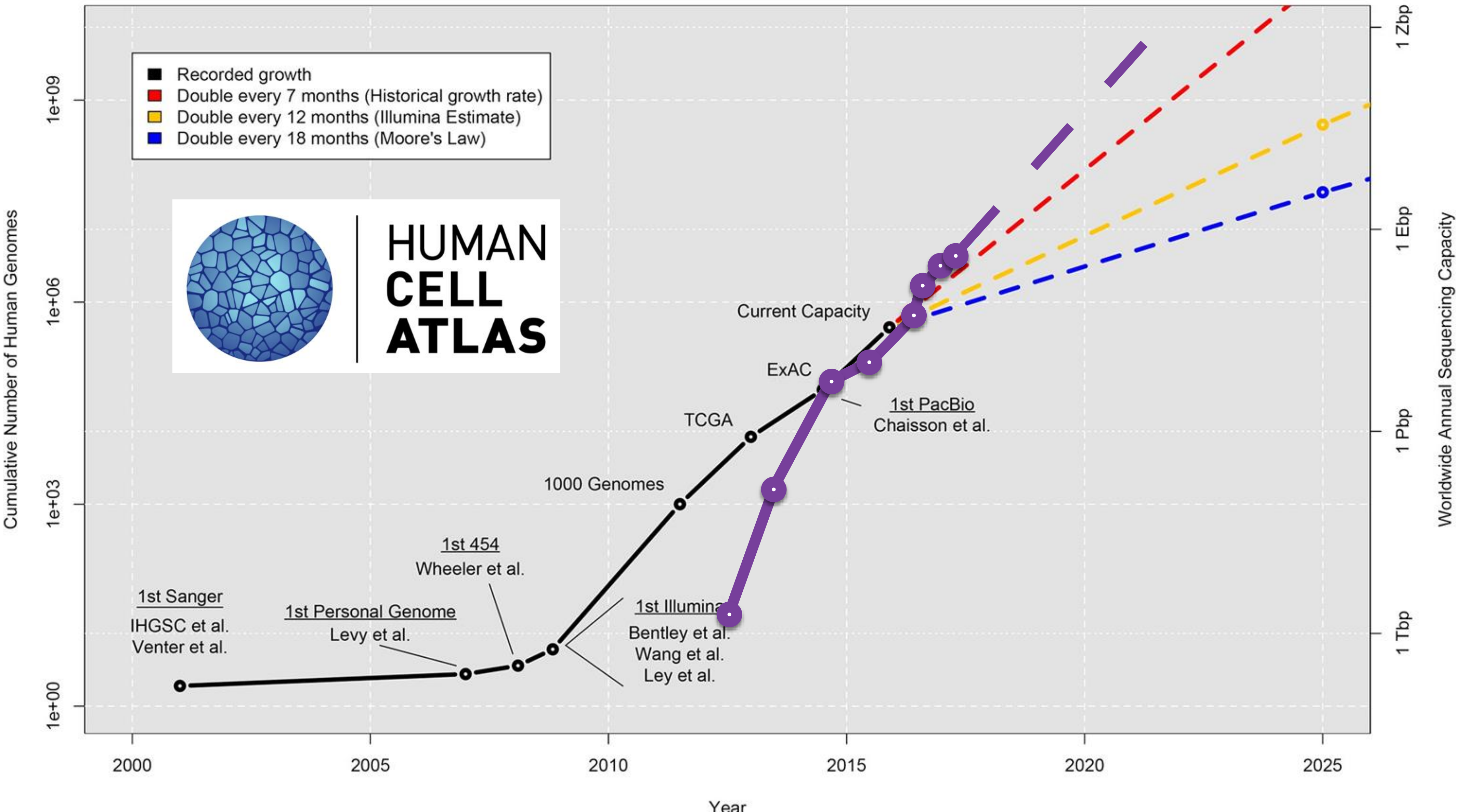
- Secure multi-party regression at plain text speed

- Addresses issues of data ownership, transmission, and storage at scale



BROAD INSTITUTE

Scaling of Cells Profiled at the Klarman Cell Observatory



Bio-Inspired AI: Representation Learning of scRNA Data

Loss Landscapes of Regularized Linear Autoencoders

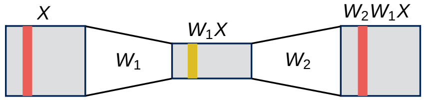
BROAD INSTITUTE[1] STANLEY CENTER[1] FOR PSYCHIATRIC RESEARCH AT BROAD INSTITUTE

Daniel Kunin*[1,2], Jonathan M. Bloom*[1], Aleksandrina Goeva[1], Cotton Seed[1]

hail.is[1] Stanford University[2]

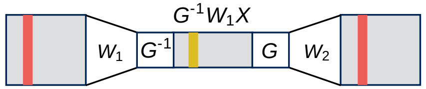
Background

A linear autoencoder maps $\mathbb{R}^m \to \mathbb{R}^k \to \mathbb{R}^m$.



$$L(W_1, W_2) = ||X - W_2W_1X||^2$$

LAEs learn the top principal *subspace* but not the principal *directions* or eigenvalues. The optimal latent representation is only defined up to a linear map $G \in GL_k(\mathbb{R})$.



LAEs are pseudoinverses at all critical pts.

Regularization

We prove that L2-*regularized* LAEs are transposes at all critical points and learn the principal directions as the left singular vectors of the decoder. Define L_σ by

$$||X - W_2W_1X||^2 + \lambda(||W_1||^2 + ||W_2||^2)$$

The minima of L_σ are defined up to an orthogonal map $O \in O_k(\mathbb{R})$ by

$$\boxed{W_2 = U_k(I - \lambda\Sigma_k^{-2})^{\frac{1}{2}}O = W_1^\intercal}$$

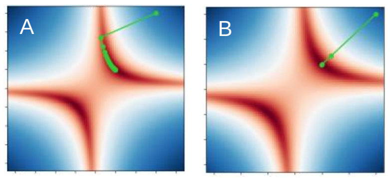
where $X = U\Sigma V^T$ and $\sigma_1^2 > \cdots > \sigma_k^2 > \lambda$.

$$\boxed{W_2W_1 = U_k(I - \lambda\Sigma_k^{-2})U_k^\intercal}$$

PCA Algorithms

Hence PCA is a two-step optimization:
1. Train L2-regularized LAE on $X \in \mathbb{R}^{m \times n}$.
2. Apply SVD to the decoder $W_2 \in \mathbb{R}^{m \times k}$.

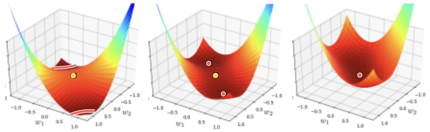
Step 2 is quick. Step 1 options include:
A. Gradient descent (below).
B. Solve for W_2, set $W_1 = W_2^\intercal$, iterate.



input $X \in \mathbb{R}^{m\times n}; k \leq m; \lambda, \alpha > 0$
initialize $W_1, W_2^\intercal \in \mathbb{R}^{k\times m}$
while not converged
  $W_1 \mathrel{-}= \alpha\,(W_2^\intercal(W_2W_1 - I)XX^\intercal + \lambda W_1)$
  $W_2 \mathrel{-}= \alpha\,((W_2W_1 - I)XX^\intercal W_1^\intercal + \lambda W_2)$
$U, \Sigma, \_ = \mathrm{SVD}(W_2)$
return $U,\ \lambda(I - \Sigma^2)^{-1}$

Posterior Collapse

Principal directions with eigenvalues below λ collapse as in probabilistic PCA.



$\lambda = 0$ $\lambda = 2$ $\lambda = 4$

Example of collapse for $X = [2]$.

Symmetry and Backprop

L2-reg LAEs are symmetric at all critical pts.

Theorem 2.1 (Transpose Theorem). *All critical points of $\mathcal{L}_\sigma$ satisfy $W_1 = W_2^\intercal$.*

Proof. Critical points of $\mathcal{L}_\sigma$ satisfy:

$$\frac{\partial \mathcal{L}_\sigma}{\partial W_1} = 2W_2^\intercal(W_2W_1 - I)XX^\intercal + 2\lambda W_1 = 0,$$
$$\frac{\partial \mathcal{L}_\sigma}{\partial W_2} = 2(W_2W_1 - I)XX^\intercal W_1^\intercal + 2\lambda W_2 = 0.$$

We first prove that the matrix

$$C = (I - W_2W_1)XX^\intercal$$

is positive semi-definite[8]. Rearranging $\frac{\partial \mathcal{L}_\sigma}{\partial W_2} W_2^\intercal$ gives

$$XX^\intercal(W_2W_1)^\intercal = (W_2W_1)XX^\intercal(W_2W_1)^\intercal + \lambda W_2W_2^\intercal.$$

Both terms on the right are positive semi-definite, so their sum on the left is as well and therefore

$$XX^\intercal(W_2W_1)^\intercal \succeq (W_2W_1)XX^\intercal(W_2W_1)^\intercal.$$

Cancelling $(W_2W_1)^\intercal$ via Lemma B.1 gives $C \succeq 0$.

We now show the difference $A = W_1 - W_2^\intercal$ is zero. Rearranging terms using the symmetry of C gives

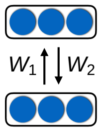
$$0 = \frac{\partial \mathcal{L}_\sigma}{\partial W_1} - \frac{\partial \mathcal{L}_\sigma}{\partial W_2}^\intercal = 2A(C + \lambda I).$$

Since $C \succeq 0$ and $\lambda > 0$ imply $C + \lambda I \succ 0$, we conclude from

$$A(C + \lambda I)A^T = 0$$

that $A = 0$. □

Resolution to weight transport problem:



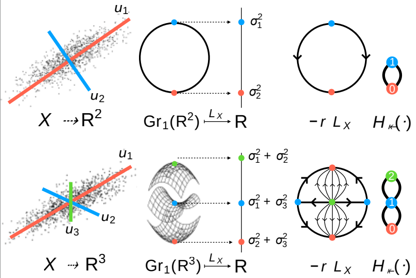
Backprop in  lacks $W_1^\intercal$ because neurons go one way.

Learn as W_2 by maximizing flow of info and minimizing energy.

Algebraic Topology

We smoothly parameterize the critical manifolds of LAEs with several forms of regularization via one elementary proof.

We factor the loss as a Morse function on the Grassmannian to reveal the dynamics near and between critical manifolds. Morse homology suggests principles and algorithms for deep learning.



Theorem 4.4 (Curvature Theorem). *In local coordinates near any point on the critical manifold indexed by $\mathcal{I}$, all three losses take the form of a standard degenerate saddle with $d_\mathcal{I} + (k - \ell)(m - \ell)$ descending directions.*

- *$\mathcal{L}$ and $\mathcal{L}_\pi$ have $k\ell$ flat directions.*
- *$\mathcal{L}_\sigma$ has $k\ell - \binom{\ell+1}{2}$ flat directions.*

The remaining directions are ascending.

Theorem E.1. *$\mathcal{L}_X$ is an $\mathbb{F}_2$-perfect Morse function. Its critical points are the rank-k principal subspaces.*

Proof. Consider the commutative diagram

$$\begin{tikzcd} \mathrm{V}_k(\mathbb{R}^m) \arrow[r, two heads, "\pi: O \mapsto \mathrm{Im}(OO^\intercal)"] \arrow[d, hook, "\iota: O \mapsto (O^\intercal, O)"'] & \mathrm{Gr}_k(\mathbb{R}^m) \arrow[d, "\mathcal{L}_X"] \\ \mathbb{R}^{k\times m} \times \mathbb{R}^{m\times k} \arrow[r, "\mathcal{L}"'] & \mathbb{R} \end{tikzcd} \quad (10)$$

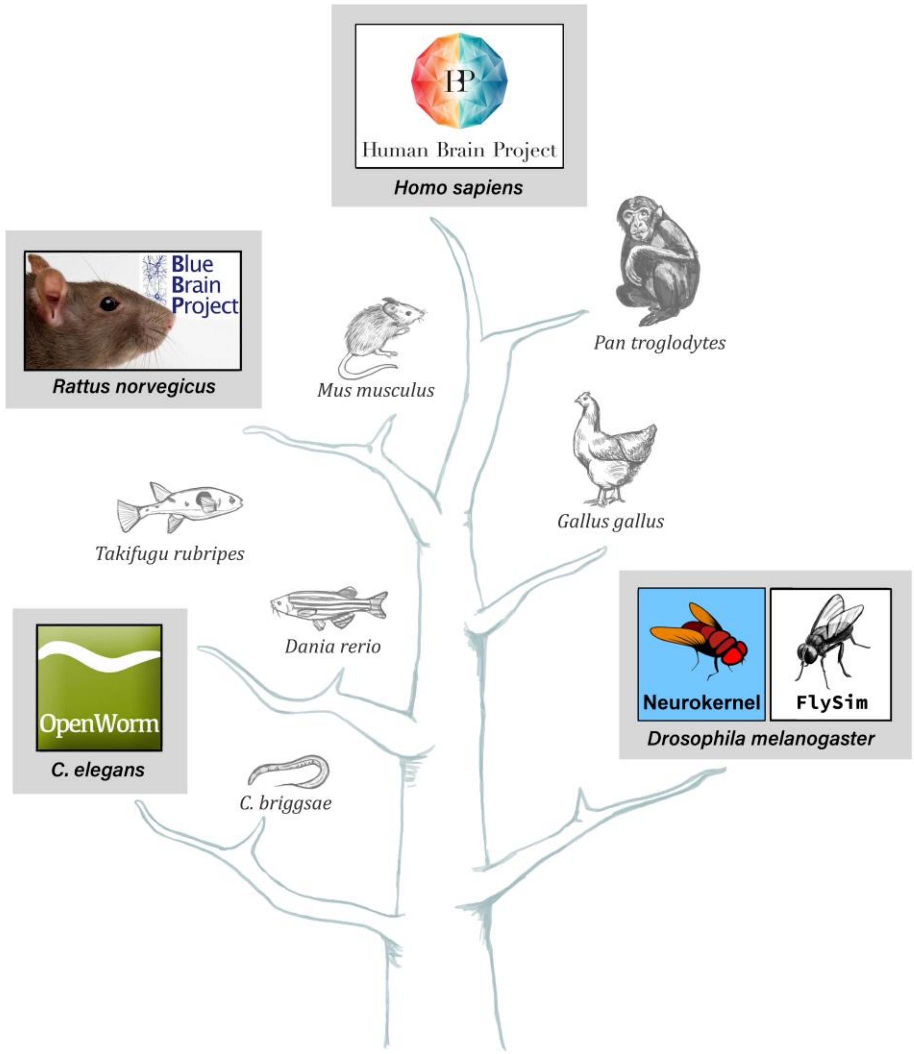
MIA talks at broadinstitute.org/mia

Bio-Inspired AI: Realistic Nervous System Simulations

G. Sarma, A. Safron, and N. Hay, “Integrative Biological Simulation, Neuropsychology, and AI Safety.” *AAAI Workshop on AI Safety* (2019)

- Simple organisms show complex behavior that continues to be difficult for modern AI systems.

- Neuronal simulations in virtual environments will allow these biological architectures to be used for AI research.



NEURON

BluePyOpt

NetPyNE

Bionet

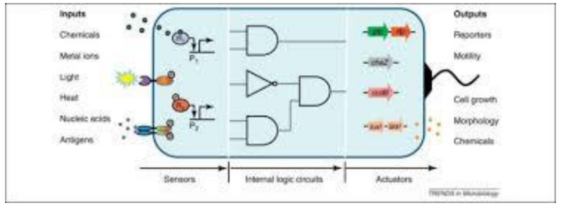
Geppetto

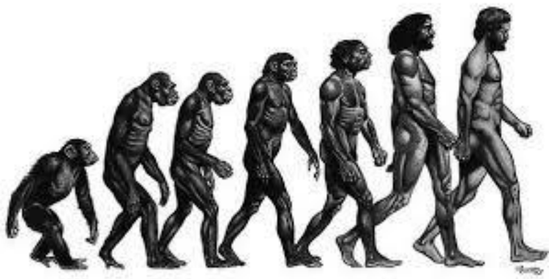
ChannelPedia

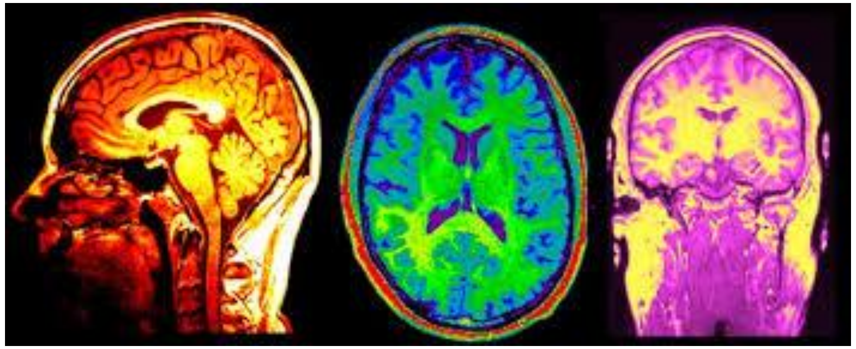
NeuroMLDB

BROAD INSTITUTE

RECURSION pharmaceuticals
flatiron Acquired by Roche
Cyft.
oscar
Google Health
Google DeepMind
Emerald Cloud Lab
BioBridge GLOBAL
insitro
Roche
BenevolentAI
verily
ROIVANT SCIENCES
HVAEN
GINKGO BIOWORKS THE ORGANISM COMPANY
PathAI
NOVARTIS
ASIMOV
Y Combinator







Complex Interplay Between AI and the Life Sciences

- **Claim 2:** Integrating AI safety and the life sciences is critical and will require significant transdisciplinary efforts

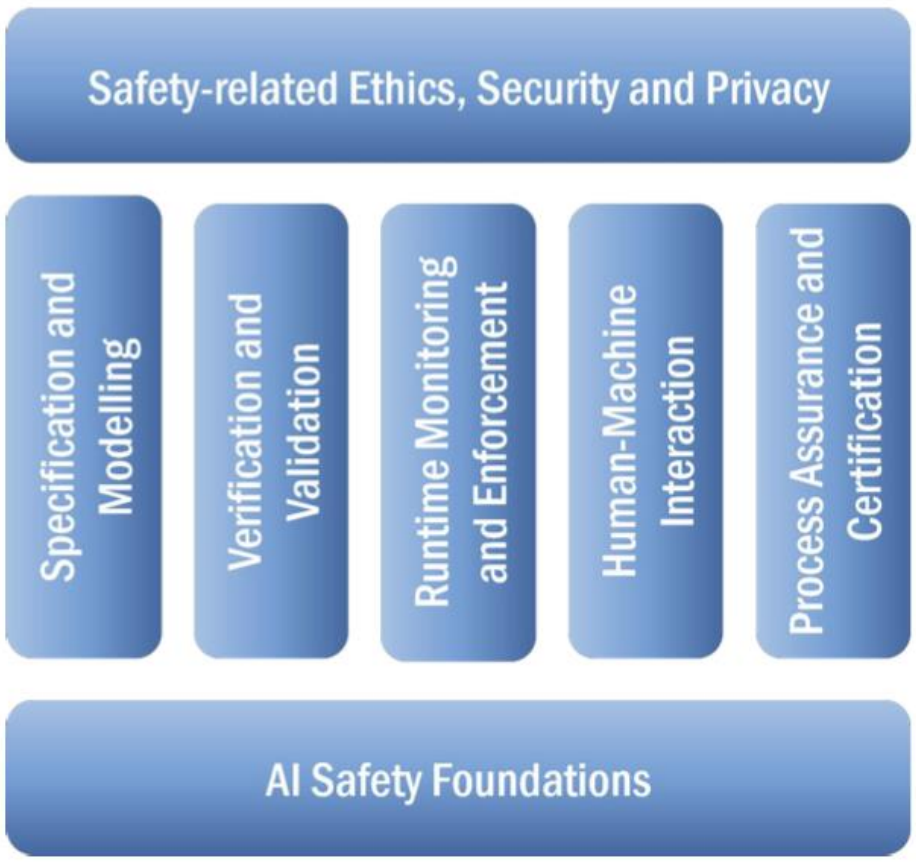
BROAD INSTITUTE

Complex Interplay Between AI and the Life Sciences

- AI will infuse genomics, cell biology, pharmaceuticals, medical devices, and care delivery
- Nearly all areas of AI safety are relevant to the life sciences
- A **culture of safety** is needed to address a diversity of safety challenges
- Biosecurity and biosafety are models for policy and cultural integration
  - 1975 Asilomar Conference on Recombinant DNA inspired Conference on Beneficial AI in 2017
  - Scientific community actively debating ethics of editing human genomes

Prioritizing the Life Sciences in the AI Safety Landscape

Life Sciences and AI Safety: A Unique Interplay of Fields



Phenome: Data security and privacy; security of medical devices; human-in-the-loop AI

System: Reproducibility of physiological models; neuroscience-inspired AGI

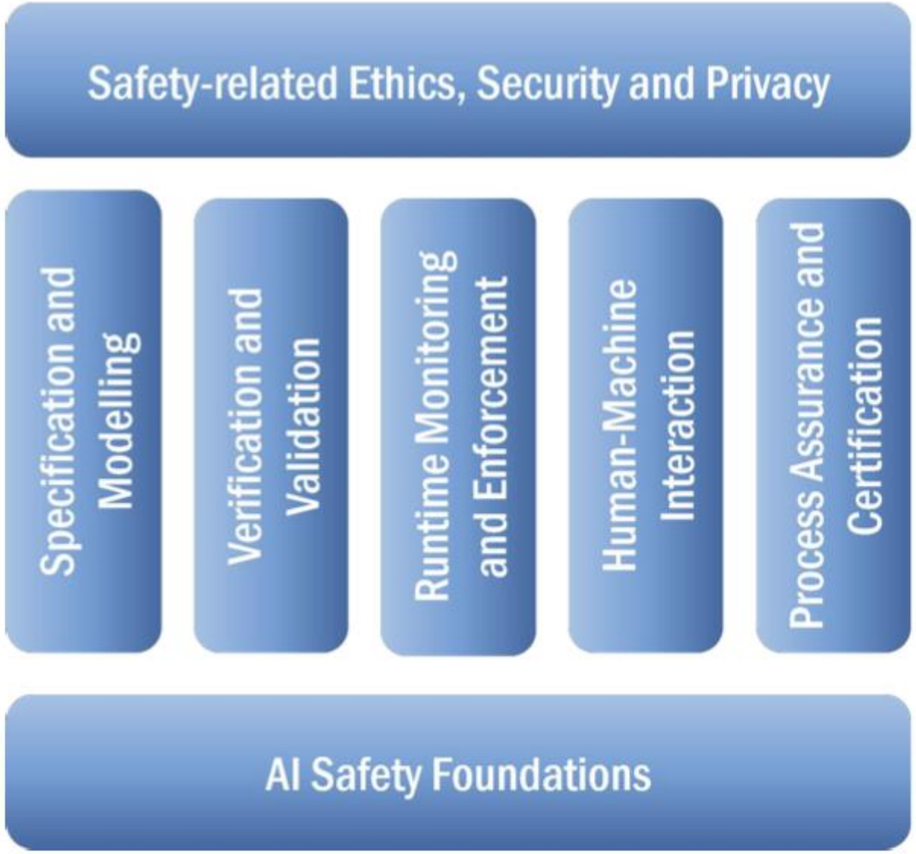
Cell: Bias in supervised learning and error propagation in medical diagnostics

Molecule: Safety and security of laboratory robotics for drug development; safe exploration

Genome: Secure multi-party computation; bias in genomic datasets

Prioritizing the Life Sciences in the AI Safety Landscape

Life Sciences and AI Safety: A Unique Interplay of Fields



Phenome: Data security and privacy; security of medical devices; human-in-the-loop AI

System: Reproducibility of physiological models; neuroscience-inspired AGI

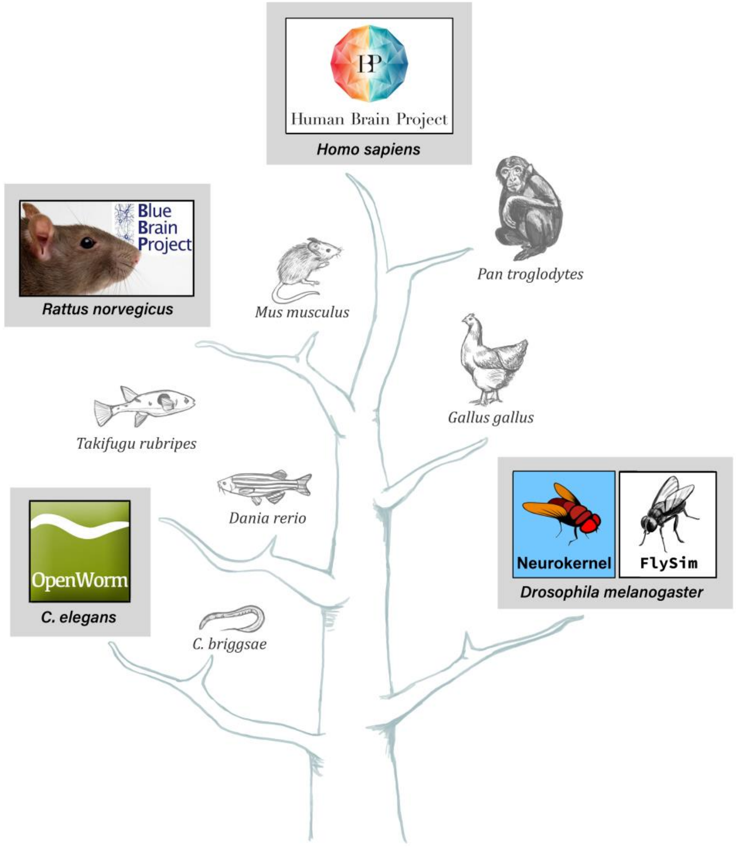
Cell: Bias in supervised learning and error propagation in medical diagnostics

Molecule: Safety and security of laboratory robotics for drug development; safe exploration

Genome: Secure multi-party computation; bias in genomic datasets

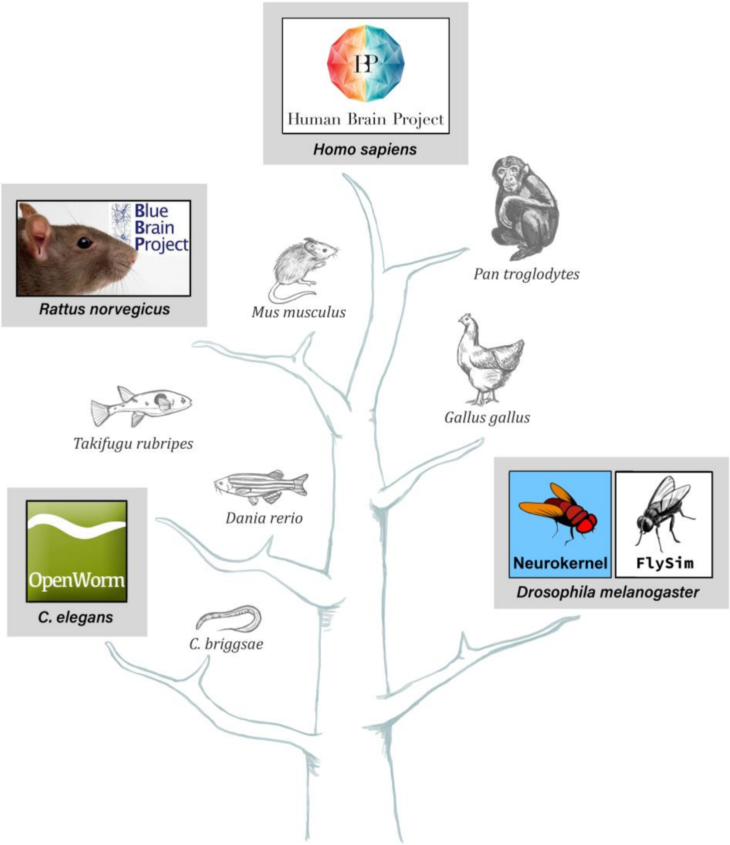
Prioritizing the Life Sciences in the AI Safety Landscape

G. Sarma, A. Safron, and N. Hay, “Integrative Biological Simulation, Neuropsychology, and AI Safety.” *AAAI Workshop on AI Safety* (2019)
J. Leike, et al. "AI Safety Gridworlds." *arXiv preprint arXiv:1711.09883* (2017).

Biological Agent and Environment Simulations	Grid Worlds Problems
	• Avoiding negative side effects • Avoiding reward hacking • Safe exploration • Safe interruptibility • Robustness to distributional shift

Prioritizing the Life Sciences in the AI Safety Landscape

G. Sarma, A. Safron, and N. Hay, “Integrative Biological Simulation, Neuropsychology, and AI Safety.” *AAAI Workshop on AI Safety* (2019)
J. Leike, et al. "AI Safety Gridworlds." *arXiv preprint arXiv:1711.09883* (2017).

<table>
<tr><th>Biological Agent and Environment Simulations</th><th>Grid Worlds Problems</th></tr>
<tr><td>


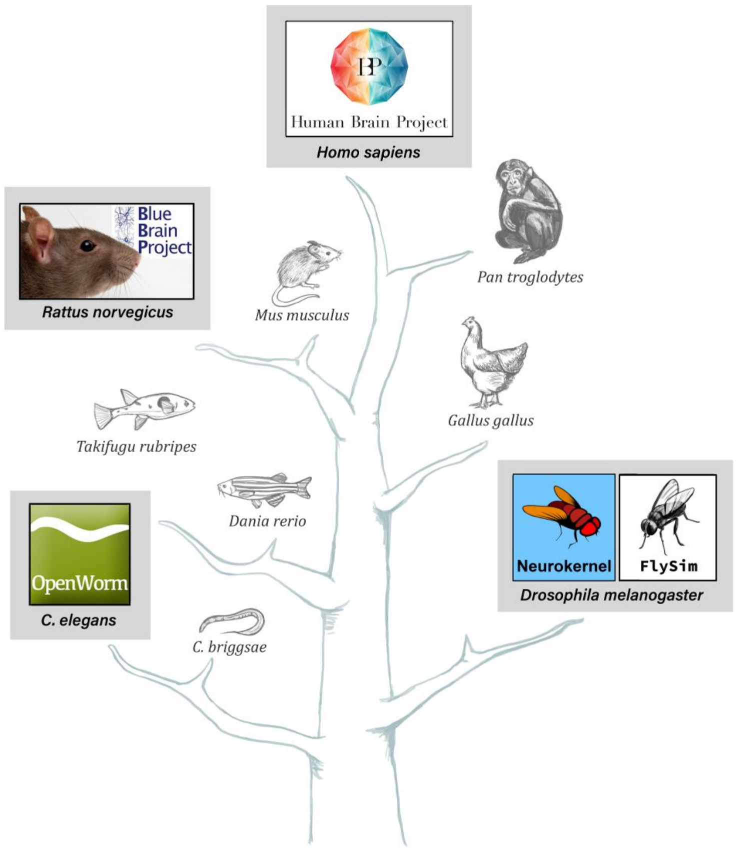
</td><td>

- Avoiding negative side effects
- **Avoiding reward hacking**
- Safe exploration
- Safe interruptibility
- Robustness to distributional shift

</td></tr>
</table>

Prioritizing the Life Sciences in the AI Safety Landscape

G. Sarma, A. Safron, and N. Hay, “Integrative Biological Simulation, Neuropsychology, and AI Safety.” *AAAI Workshop on AI Safety* (2019)
J. Leike, et al. "AI Safety Gridworlds." *arXiv preprint arXiv:1711.09883* (2017).

Biological Agent and Environment Simulations	Grid Worlds Problems
	• Avoiding negative side effects • Avoiding reward hacking • **Safe exploration** • Safe interruptibility • Robustness to distributional shift

Summary

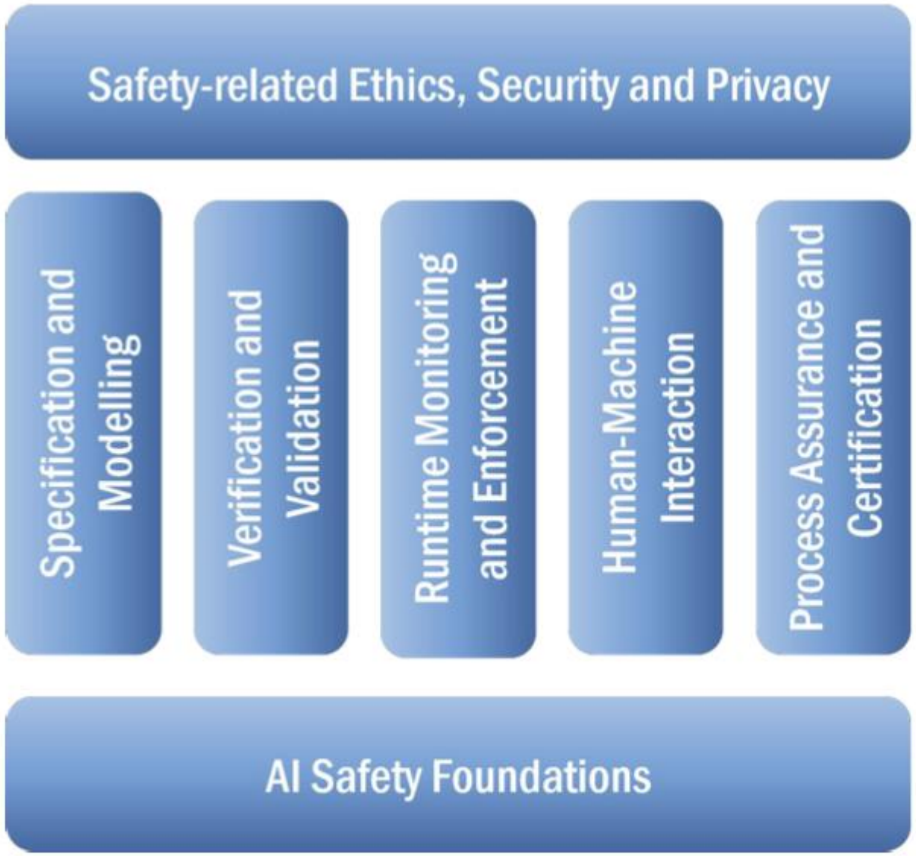
- **Claim 1:** Bio-inspired AI: life sciences will advance AI itself

- **Claim 2:** Bio-inspired AI safety: the life sciences will advance AI safety itself

BROAD INSTITUTE

Summary

Life Sciences and AI Safety: A Unique Interplay of Fields



Phenome: Data security and privacy; security of medical devices; human-in-the-loop AI

System: Reproducibility of physiological models; neuroscience-inspired AGI

Cell: Bias in supervised learning and error propagation in medical diagnostics

Molecule: Safety and security of laboratory robotics for drug development; safe exploration

Genome: Secure multi-party computation; bias in genomic datasets

https://broadinstitute.org/mia

Follow @gopalpsarma

gsarma@broadinstitute.org