DIGITAL TWINS : CRASH OF DATA, MATHEMATICS, SIMULATION, MACHINE LEARNING AND VISUALIZATION

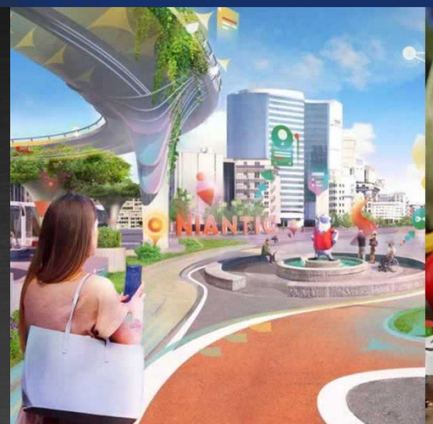
Prof (Dr) Simon See
Head of AI Technology Center, NVIDIA
Prof (Adj), Mahindra, UI, KMUTT, BUPT, SJTU, NTU
Distinguish Fellow, Fudan Uni





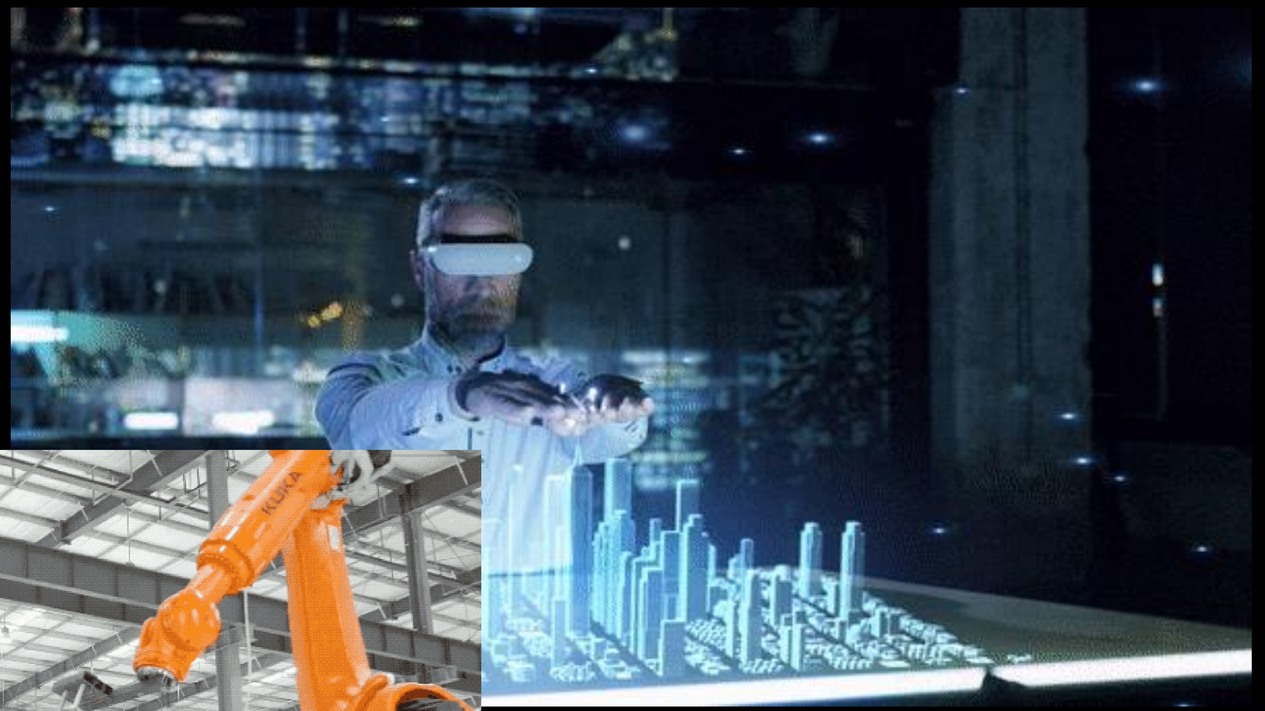


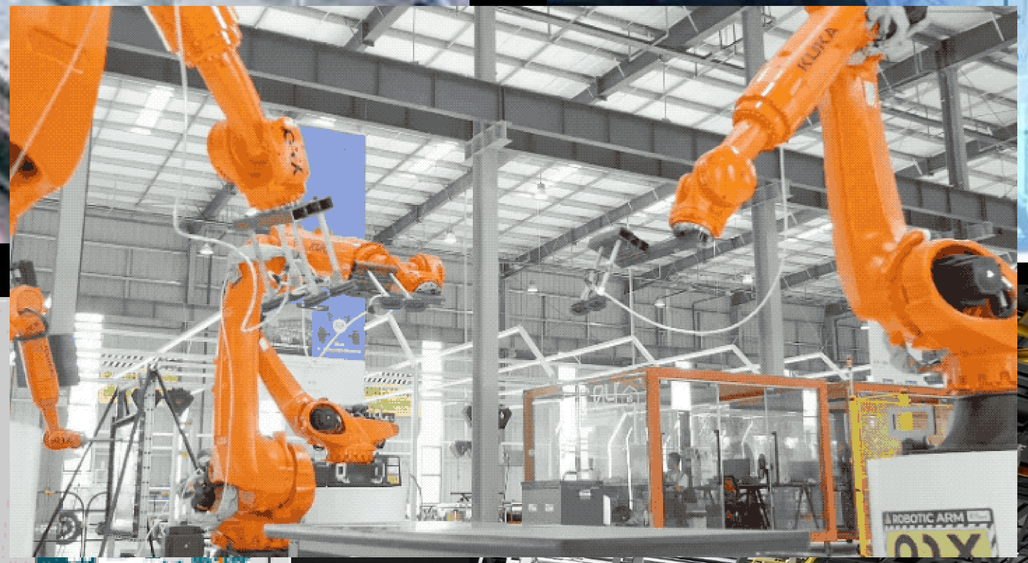






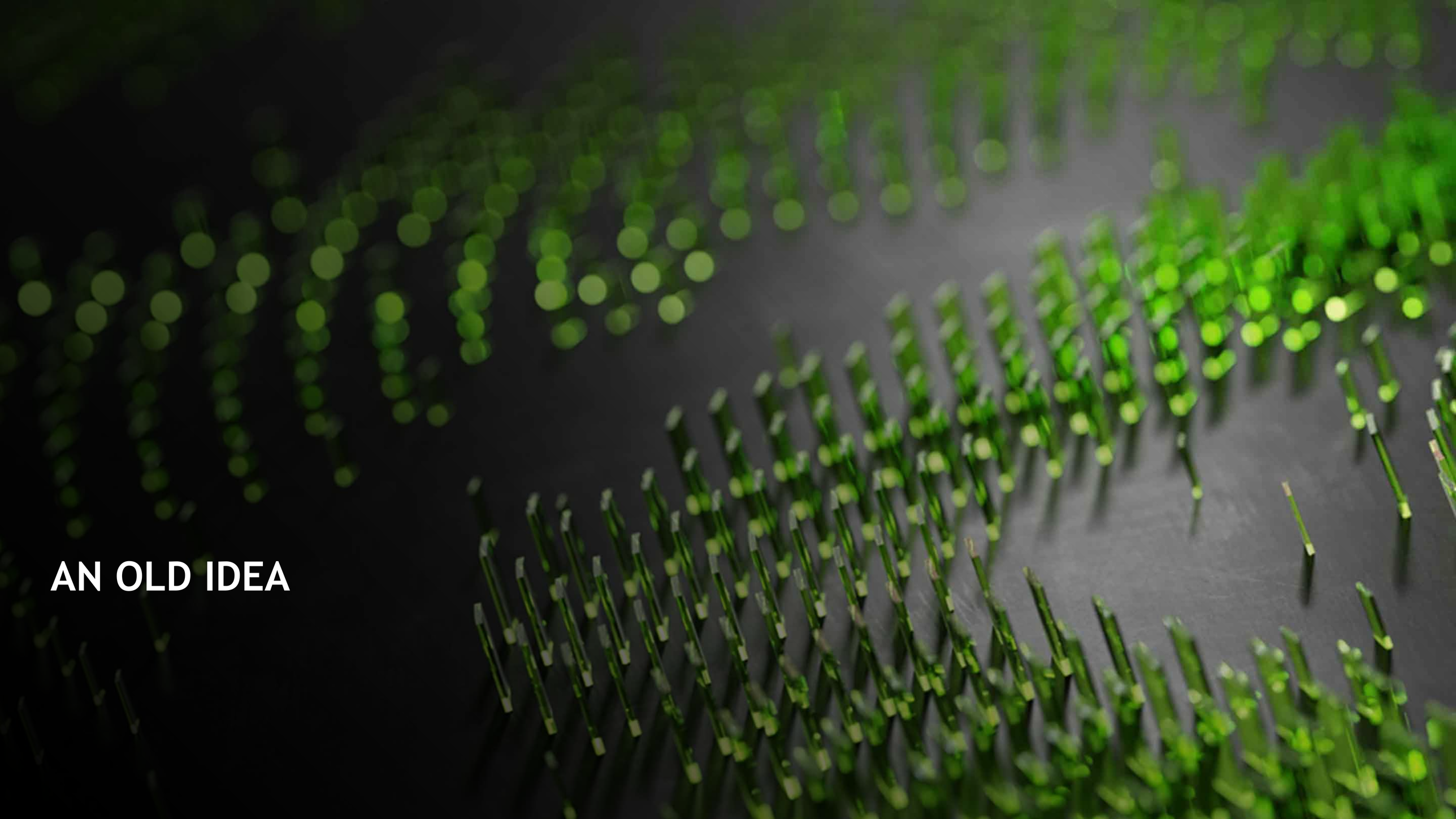


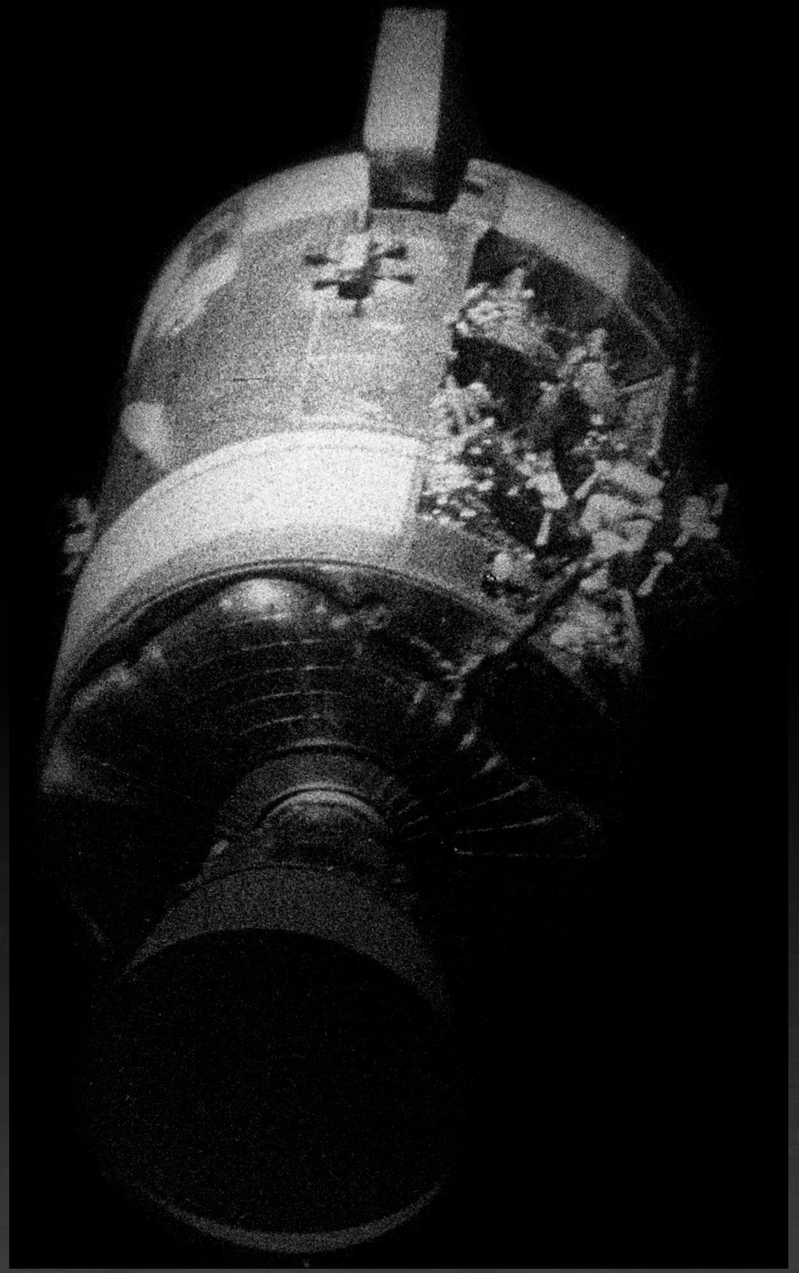

KUKA



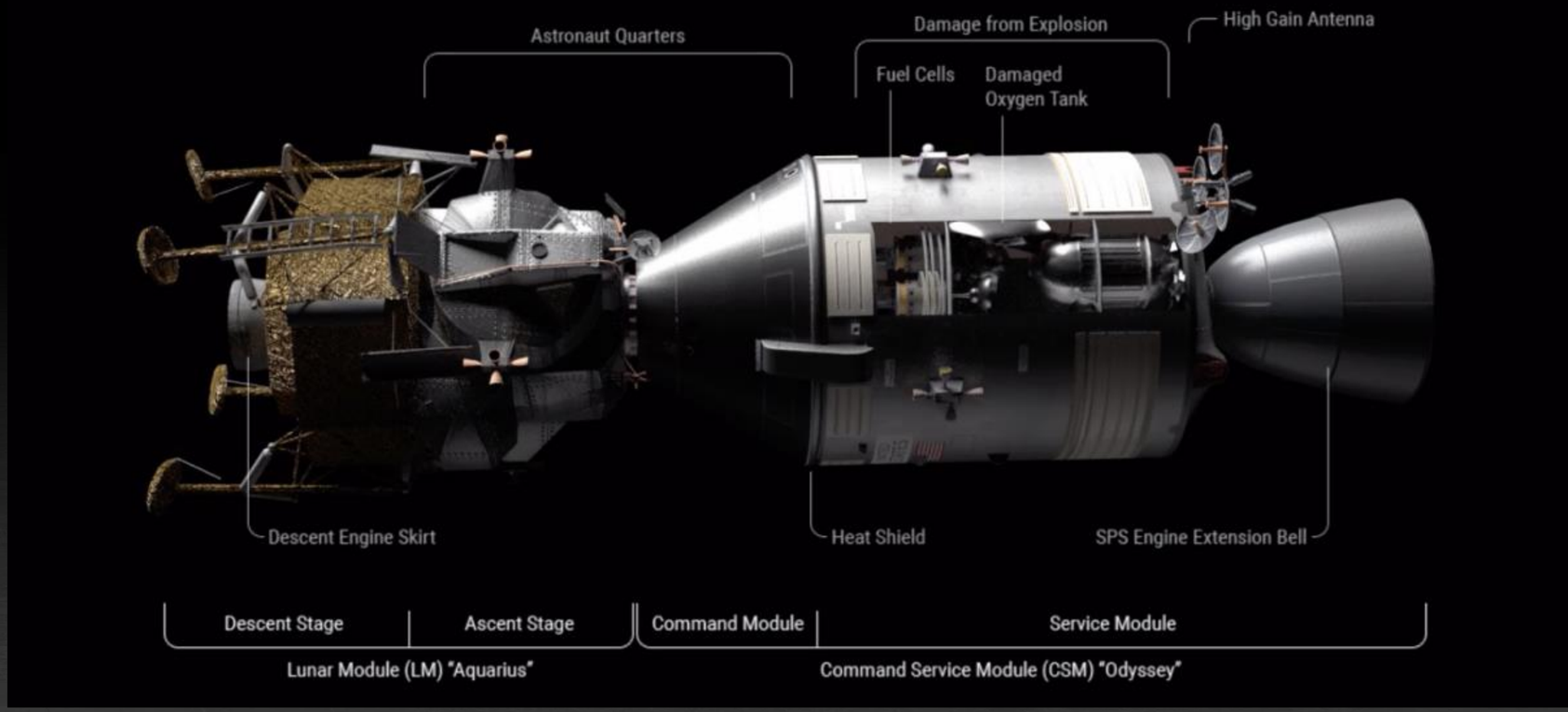


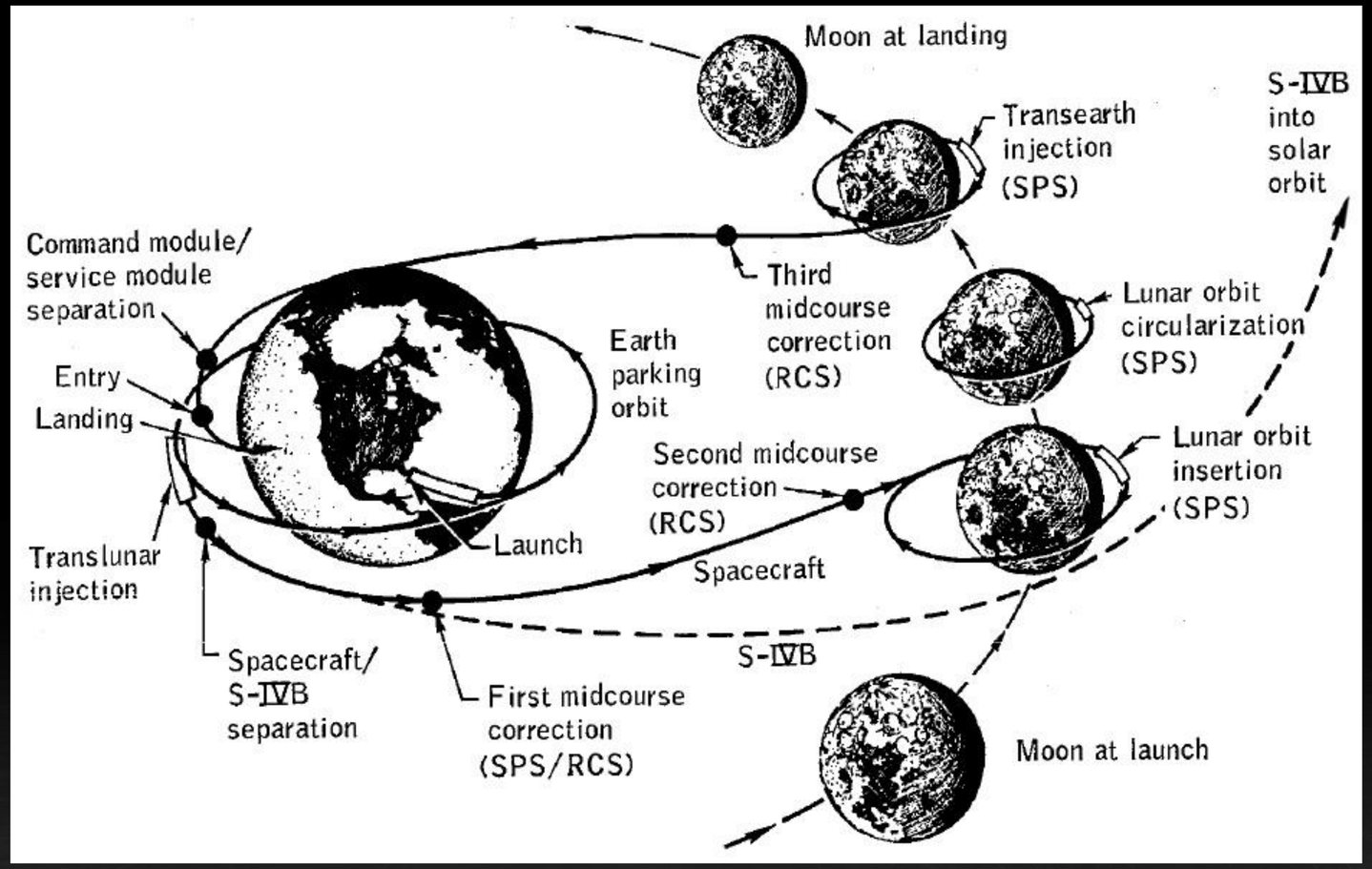
AN OLD IDEA





APOLLO 13

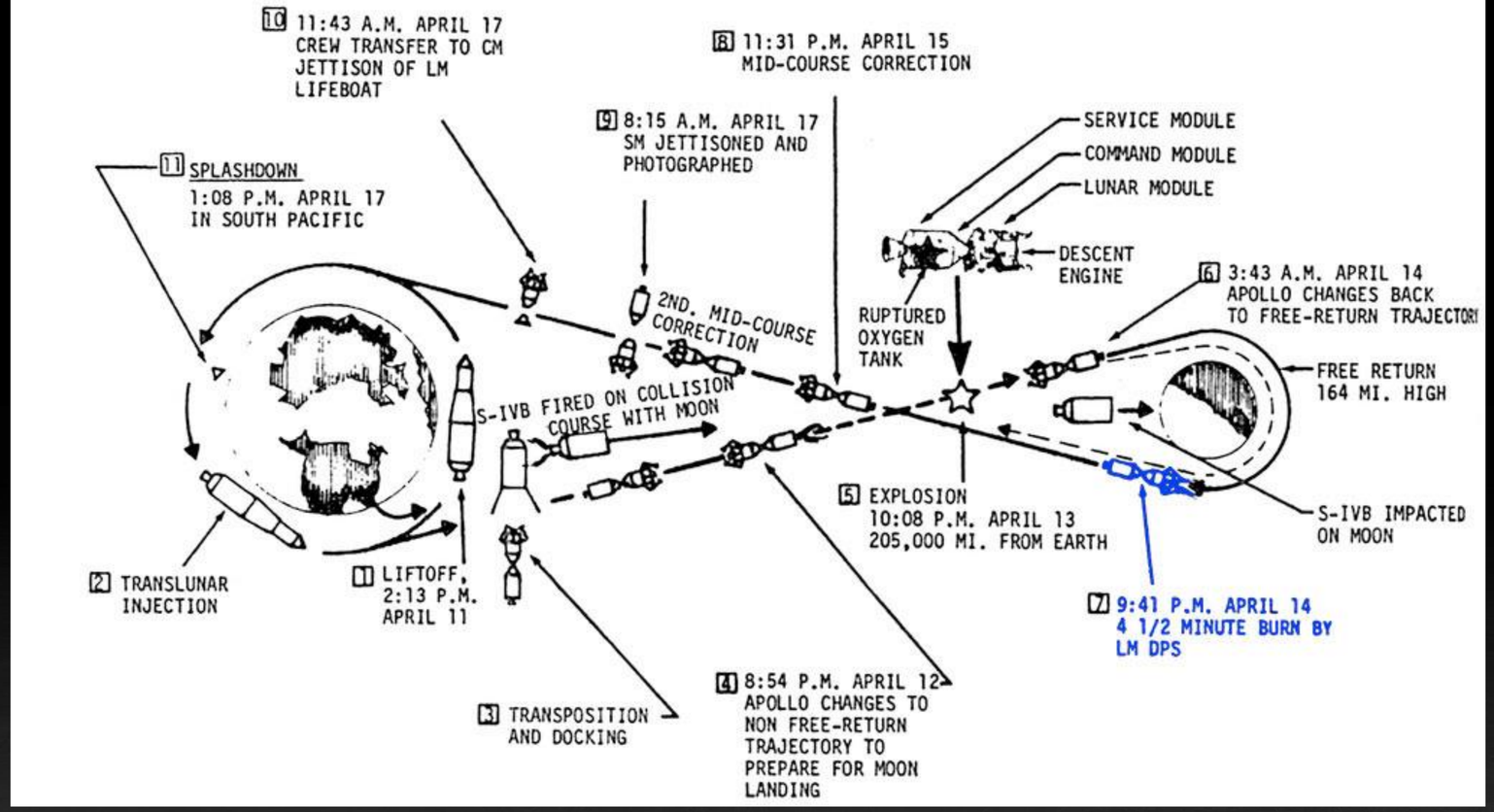

Astronaut Quarters
Damage from Explosion
High Gain Antenna
Fuel Cells
Damaged
Oxygen Tank
Descent Engine Skirt
Heat Shield
SPS Engine Extension Bell
Descent Stage
Ascent Stage
Command Module
Service Module
Lunar Module (LM) "Aquarius"
Command Service Module (CSM) "Odyssey"

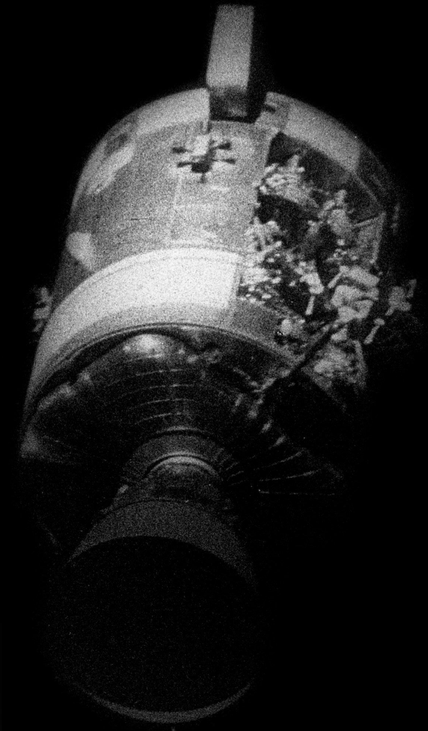

Moon at landing
Transearth injection (SPS)
S-IVB into solar orbit
Command module/ service module separation
Third midcourse correction (RCS)
Lunar orbit circularization (SPS)
Earth parking orbit
Entry
Landing
Second midcourse correction (RCS)
Lunar orbit insertion (SPS)
Launch
Translunar injection
Spacecraft
S-IVB
Spacecraft/ S-IVB separation
First midcourse correction (SPS/RCS)
Moon at launch



Apollo 13

APOLLO 13 FLIGHT PROFILE

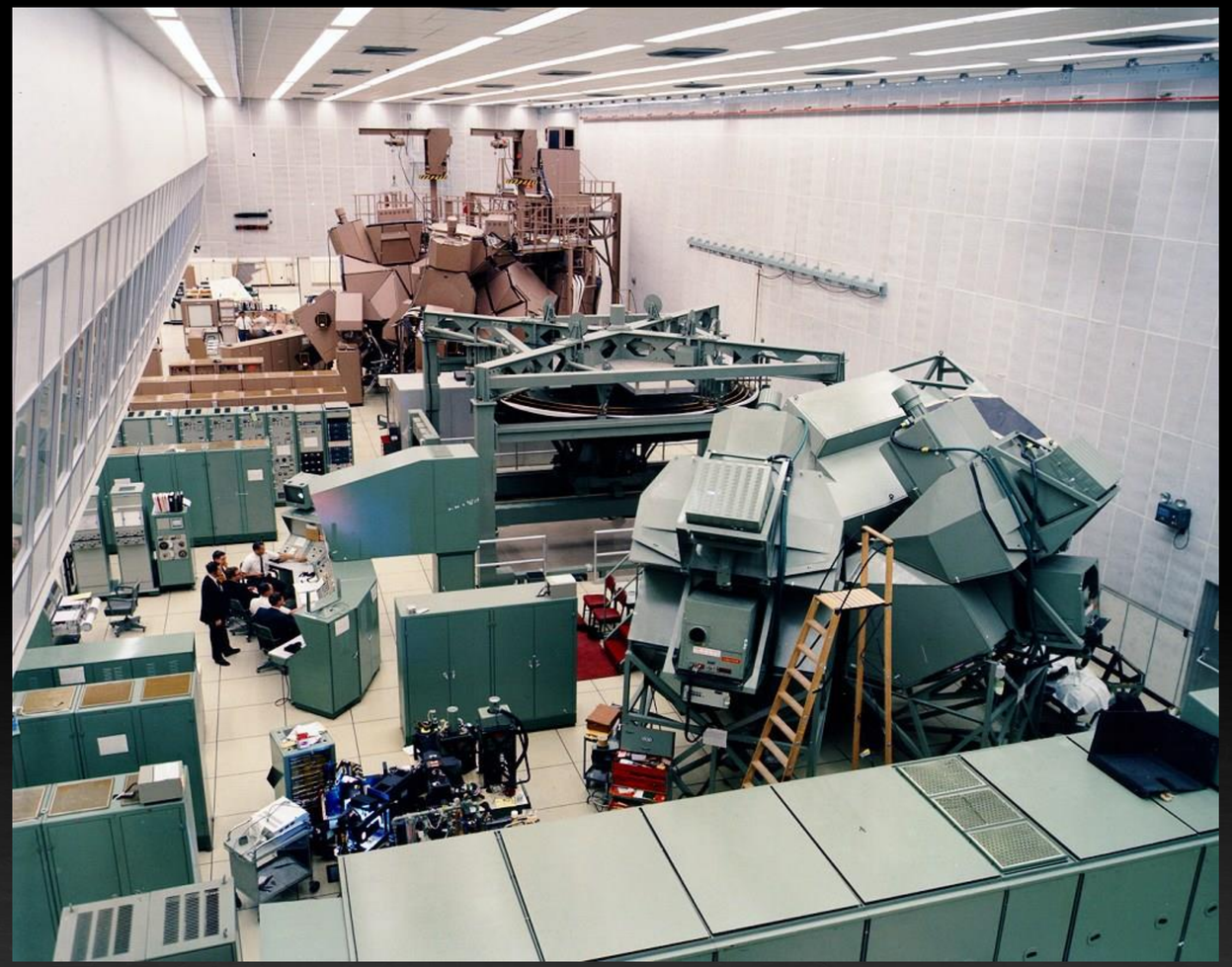






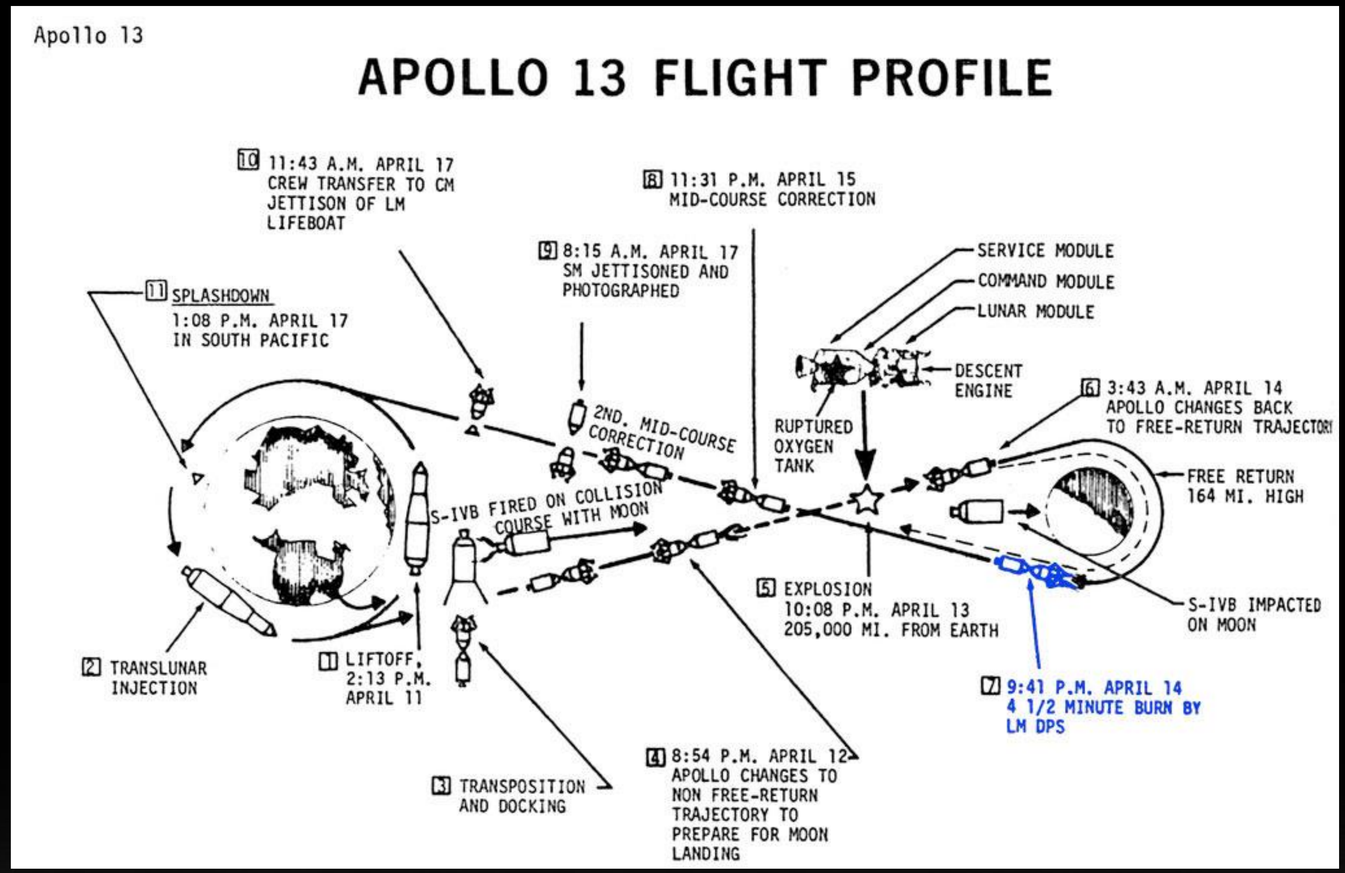
ANALOG TWINS



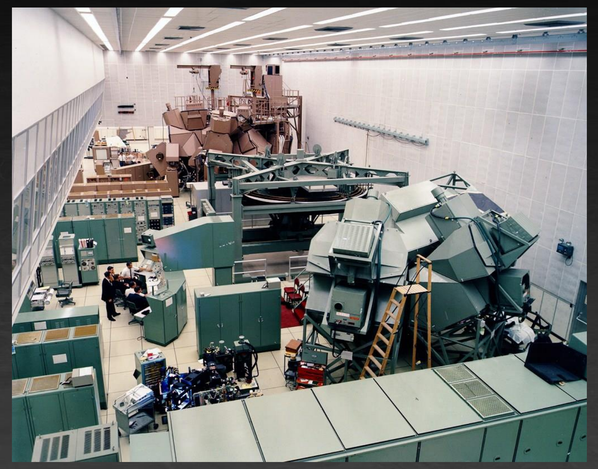


ANALOG TWINS



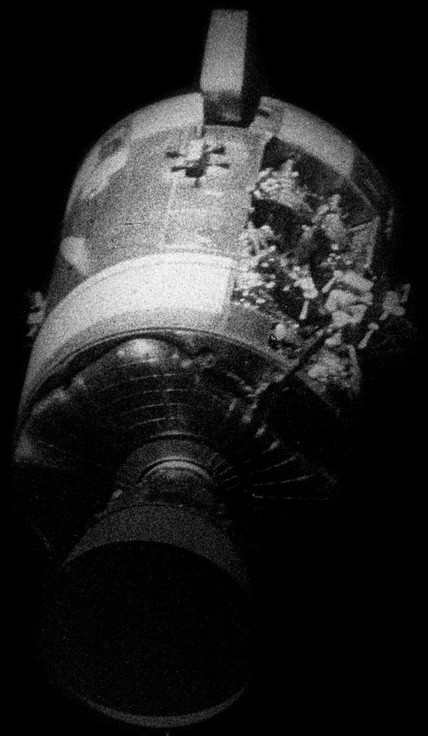


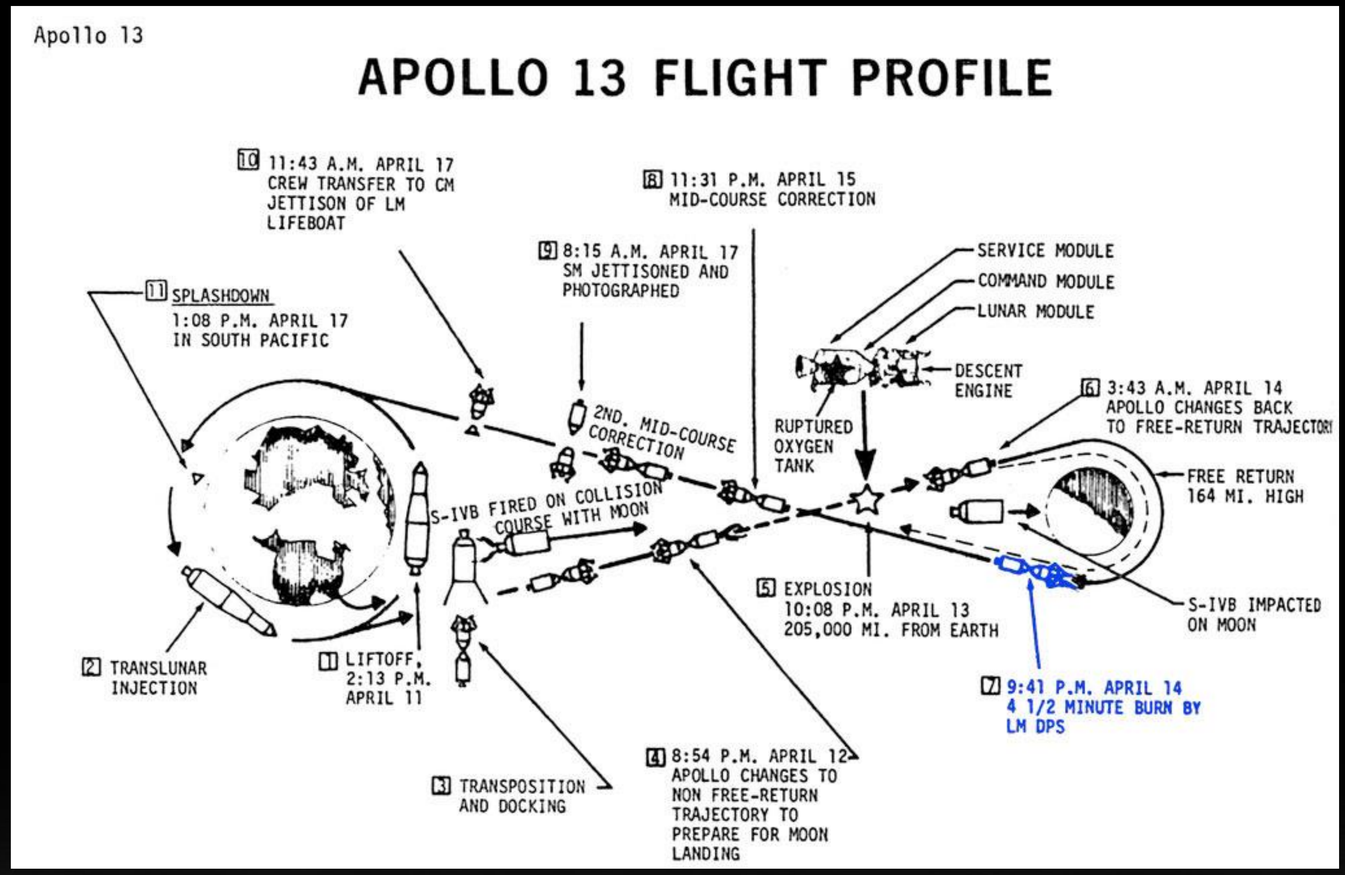




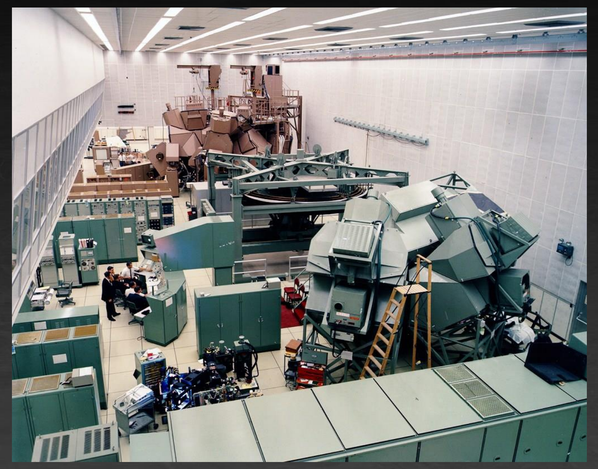


ANALOG TWINS











HARDWARE IN THE LOOP (HWIL)

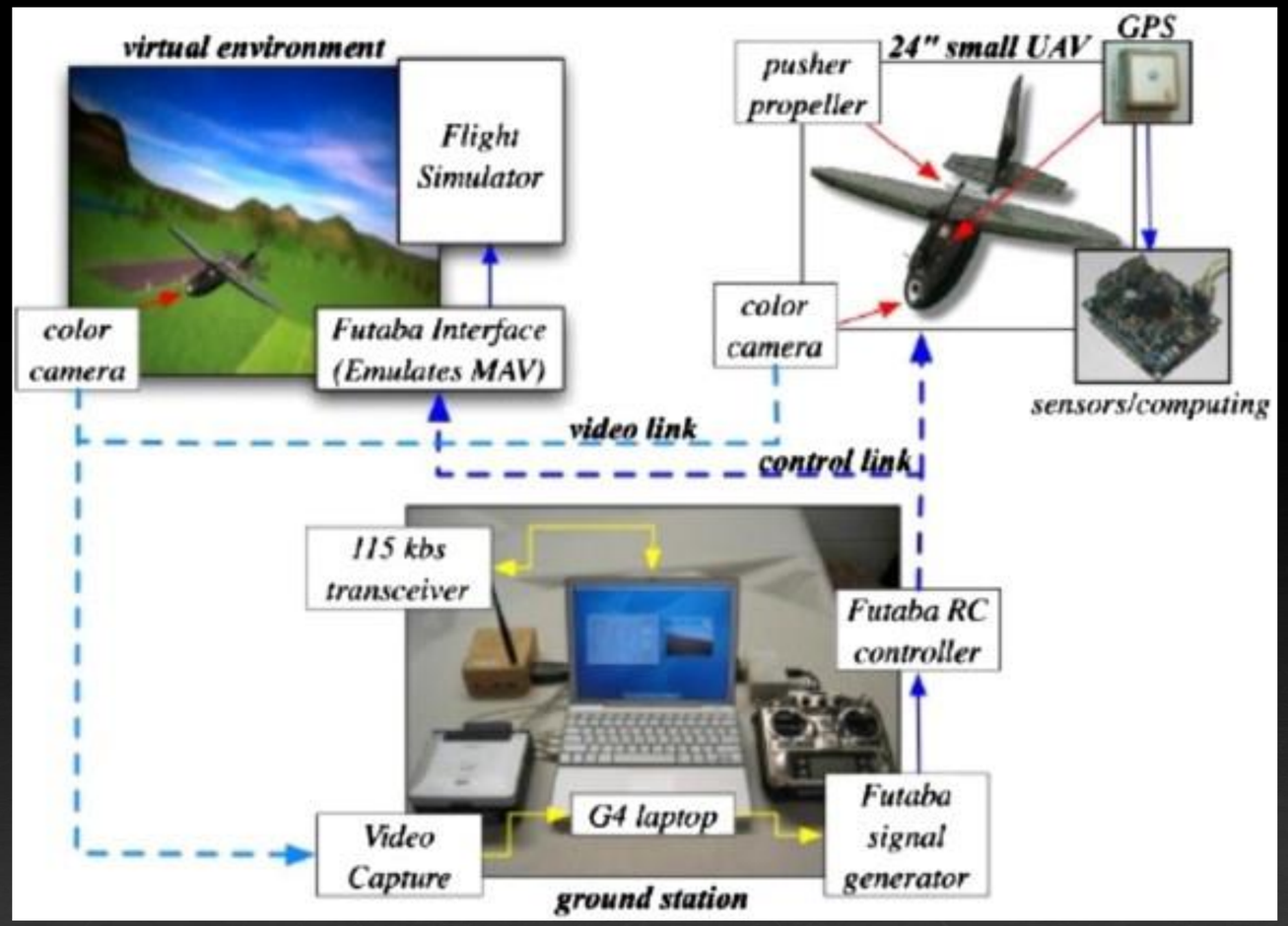


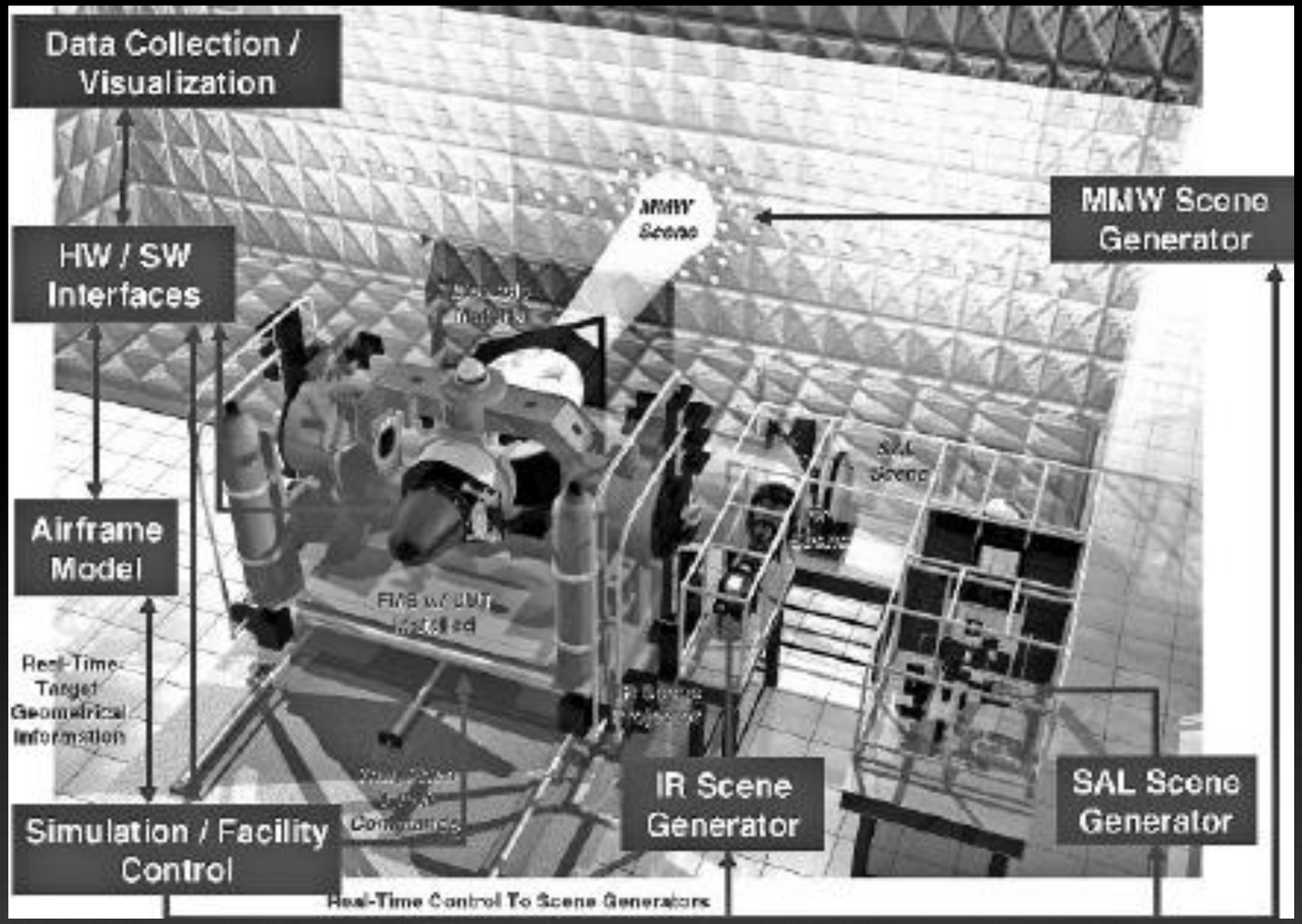




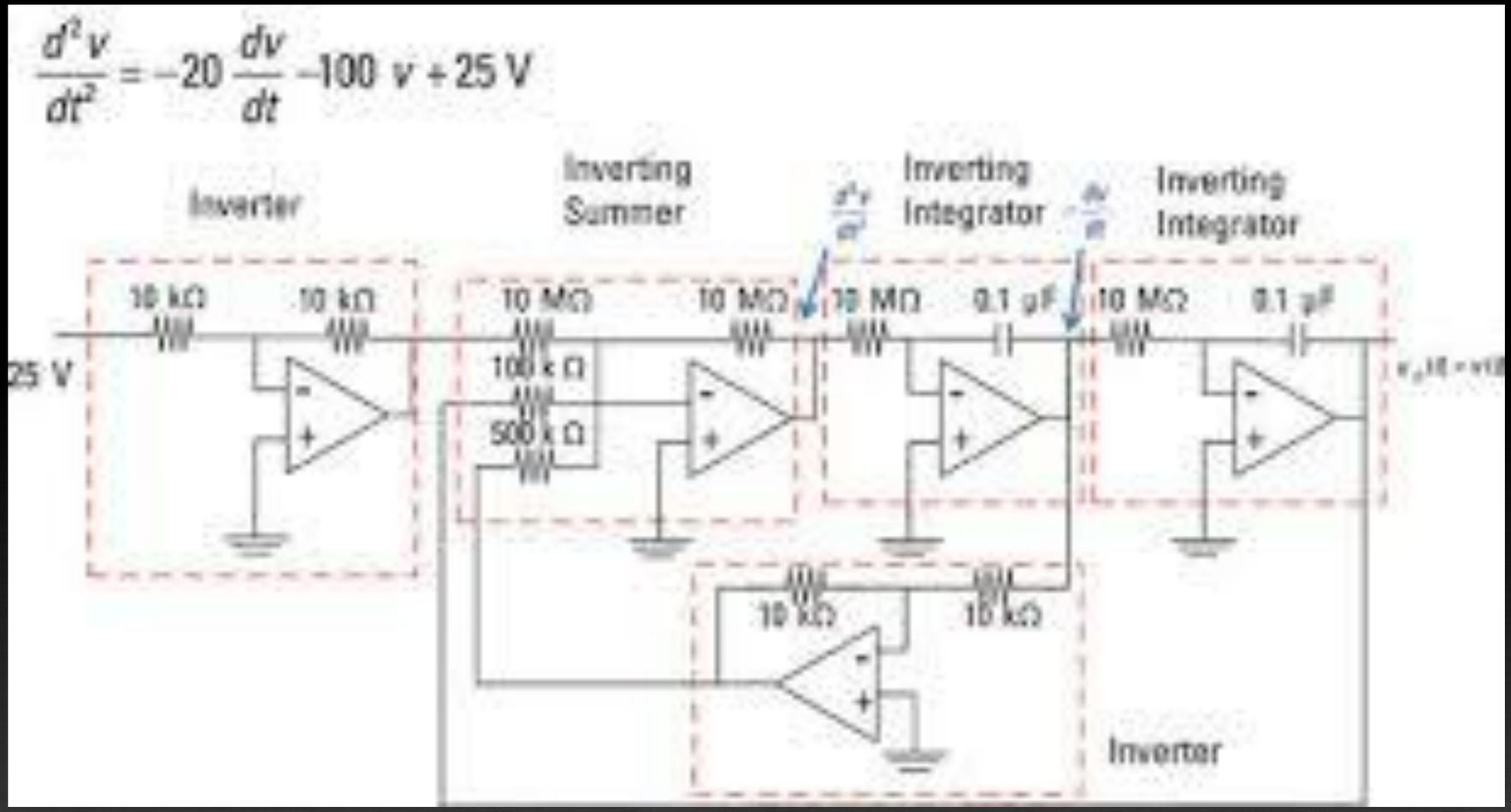
Fig. 9 Hardware-In-the-Loop simulation test

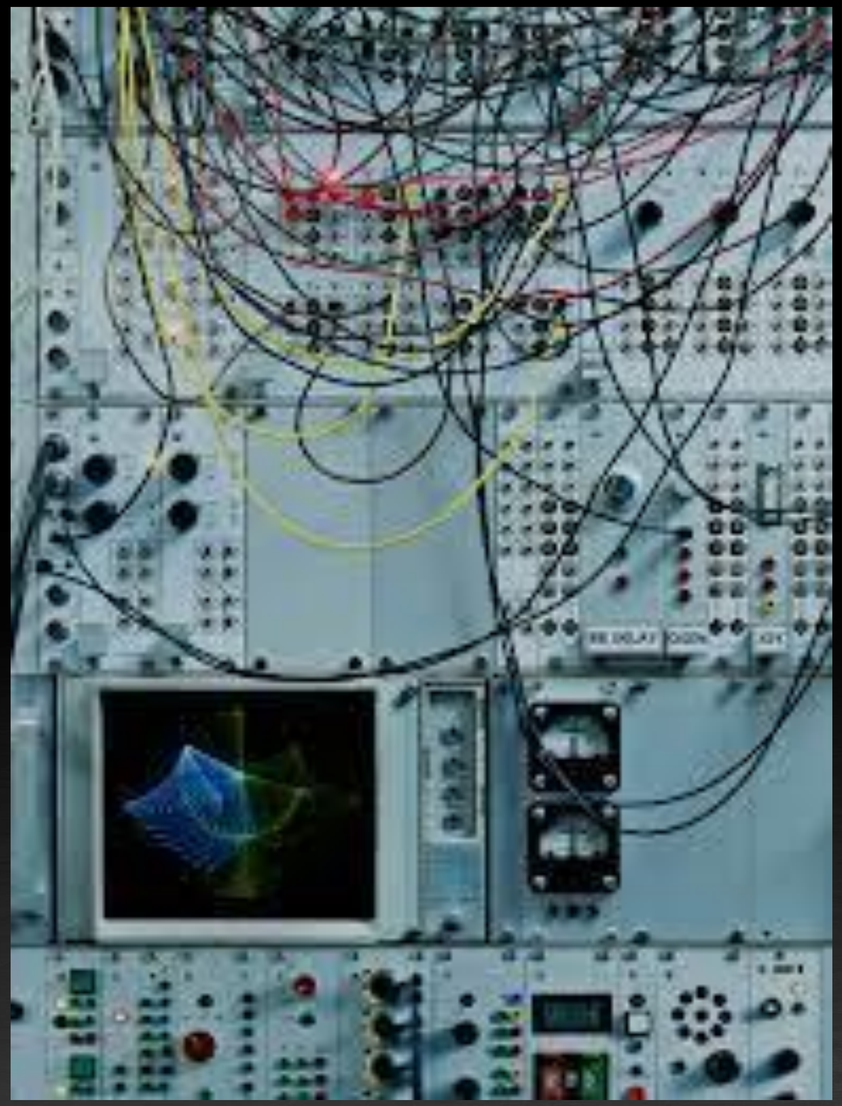


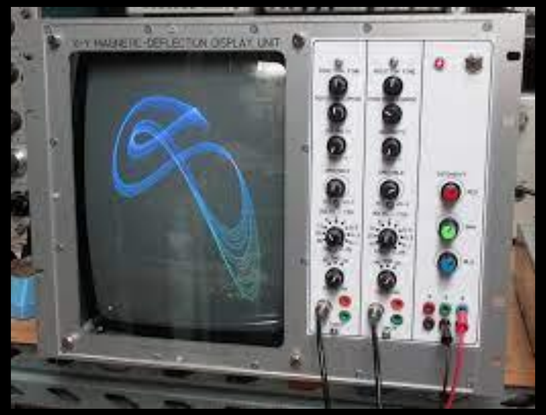



Data Collection / Visualization
HW / SW Interfaces
Airframe Model
Real-Time Target Geometrical Information
Simulation / Facility Control
MMW Scene Generator
IR Scene Generator
SAL Scene Generator
Real-Time Control To Scene Generators

Solving an Ordinary Different Equation : The Analog Way






X-Y MAGNETIC-DEFLECTION DISPLAY UNIT





FACEBOOK (META). World largest AI SuperCluster



4000+ GPU Super Cluster

REDUCED ORDER AND SURROGATE MODEL

Equations (3.a.23) – (3.a.32) become:

$$\frac{\partial \zeta}{\partial t} = \mathbf{k} \cdot \nabla \times (\mathfrak{n}/\cos\phi) + F_{\zeta_M}\,, \tag{3.a.33}$$

$$\frac{\partial \delta}{\partial t} = \nabla \cdot (\mathfrak{n}/\cos\phi) - \nabla^2 (E + \Phi) + F_{\delta_M}\,, \tag{3.a.34}$$

$$\frac{\partial T}{\partial t} = \frac{-1}{a\cos^2\phi}\left[\frac{\partial}{\partial\lambda}(UT) + \cos\phi\frac{\partial}{\partial\phi}(VT)\right] + T\delta - \dot{\eta}\frac{\partial p}{\partial\eta}\frac{\partial T}{\partial p} + \frac{R}{c_p^*}\mathcal{T}\frac{\omega}{p}$$
$$+ Q + F_{T_M} + F_{F_M}\,, \tag{3.a.35}$$

$$\frac{\partial q}{\partial t} = \frac{-1}{a\cos^2\phi}\left[\frac{\partial}{\partial\lambda}(Uq) + \cos\phi\frac{\partial}{\partial\phi}(Vq)\right] + q\delta - \dot{\eta}\frac{\partial p}{\partial\eta}\frac{\partial q}{\partial p} + S\,, \tag{3.a.36}$$

$$\frac{\partial \Pi}{\partial t} = -\int_{(\eta_k)}^{(1)} \mathbf{V}\cdot\nabla\Pi d\left(\frac{\partial p}{\partial\pi}\right) - \frac{1}{\pi}\int_{p(\eta_k)}^{p(1)} \delta dp, \tag{3.a.37}$$

$$\dot{\eta}_U = +(\zeta + f)V - \dot{\eta}\frac{\partial p}{\partial\eta}\frac{\partial U}{\partial p} - R\frac{\mathcal{T}}{a}\frac{\pi}{p}\frac{\partial p}{\partial\pi}\frac{\partial\Pi}{\partial\lambda} + F_U, \tag{3.a.38}$$

$$\dot{\eta}_V = -(\zeta + f)U - \dot{\eta}\frac{\partial p}{\partial\eta}\frac{\partial V}{\partial p} - R\frac{\mathcal{T}\cos\phi}{a}\frac{\pi}{p}\frac{\partial p}{\partial\pi}\frac{\partial\Pi}{\partial\phi} + F_V, \tag{3.a.39}$$

$$\Phi = \Phi_s + R\int_{p(\eta)}^{p(1)} \mathcal{T}\, d\ln p, \tag{3.a.40}$$

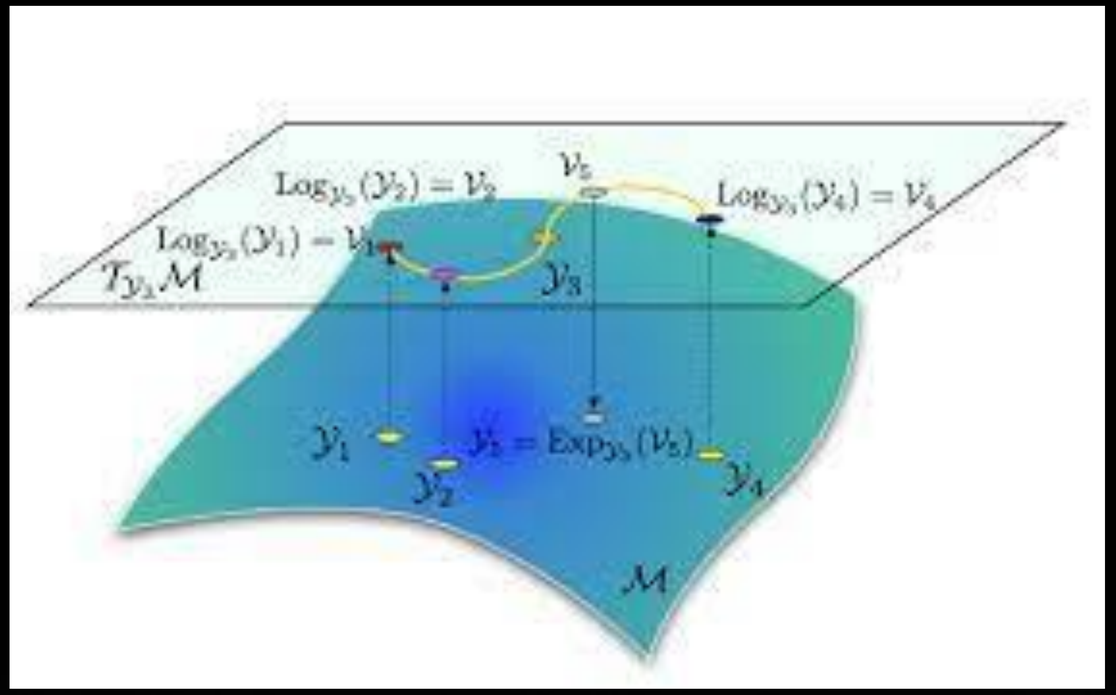
$$\dot{\eta}\frac{\partial p}{\partial\eta} = \frac{\partial p}{\partial\pi}\left[\int_{(\eta_k)}^{(1)} \pi\mathbf{V}\cdot\nabla\Pi d\left(\frac{\partial p}{\partial\pi}\right) + \int_{p(\eta_k)}^{p(1)} \delta dp\right]$$
$$- \int_{(\eta_k)}^{(\eta)} \pi\mathbf{V}\cdot\nabla\Pi d\left(\frac{\partial p}{\partial\pi}\right) - \int_{p(\eta_k)}^{p(\eta)} \delta dp, \tag{3.a.41}$$

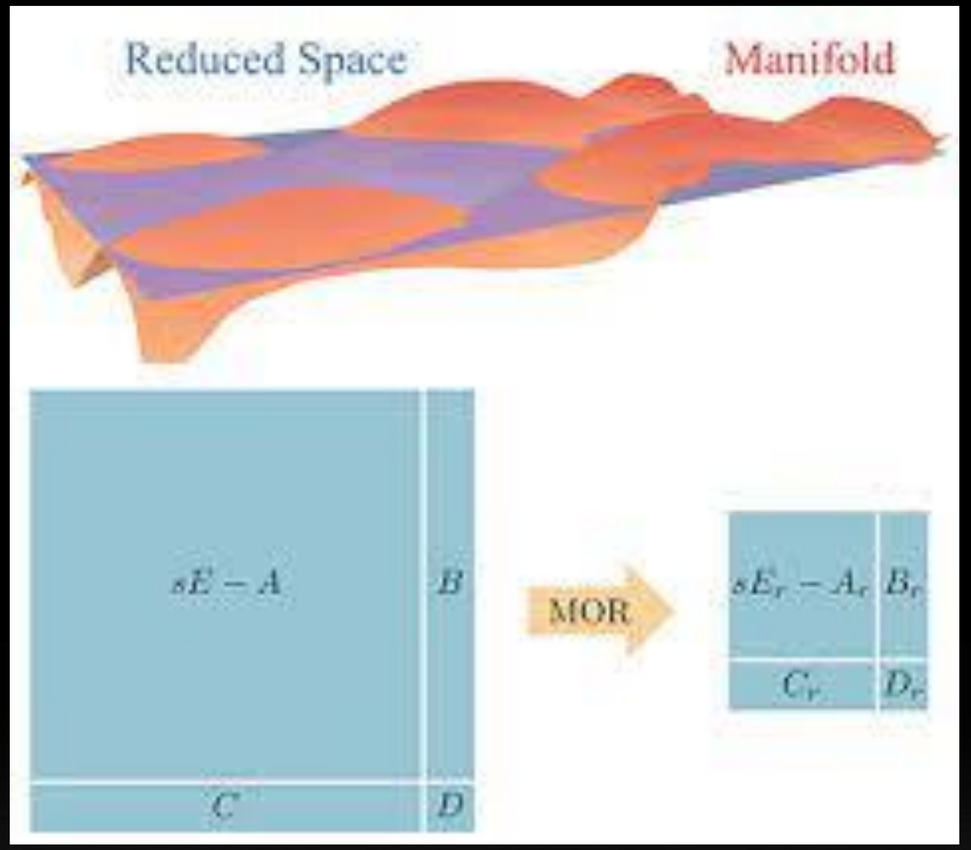
$$\omega = \frac{\partial p}{\partial\pi}\pi\mathbf{V}\cdot\nabla\Pi - \int_{(\eta_k)}^{(\eta)} \pi\mathbf{V}\cdot\nabla\Pi d\left(\frac{\partial p}{\partial\pi}\right) - \int_{p(\eta_k)}^{p(\eta)} \delta dp. \tag{3.a.42}$$







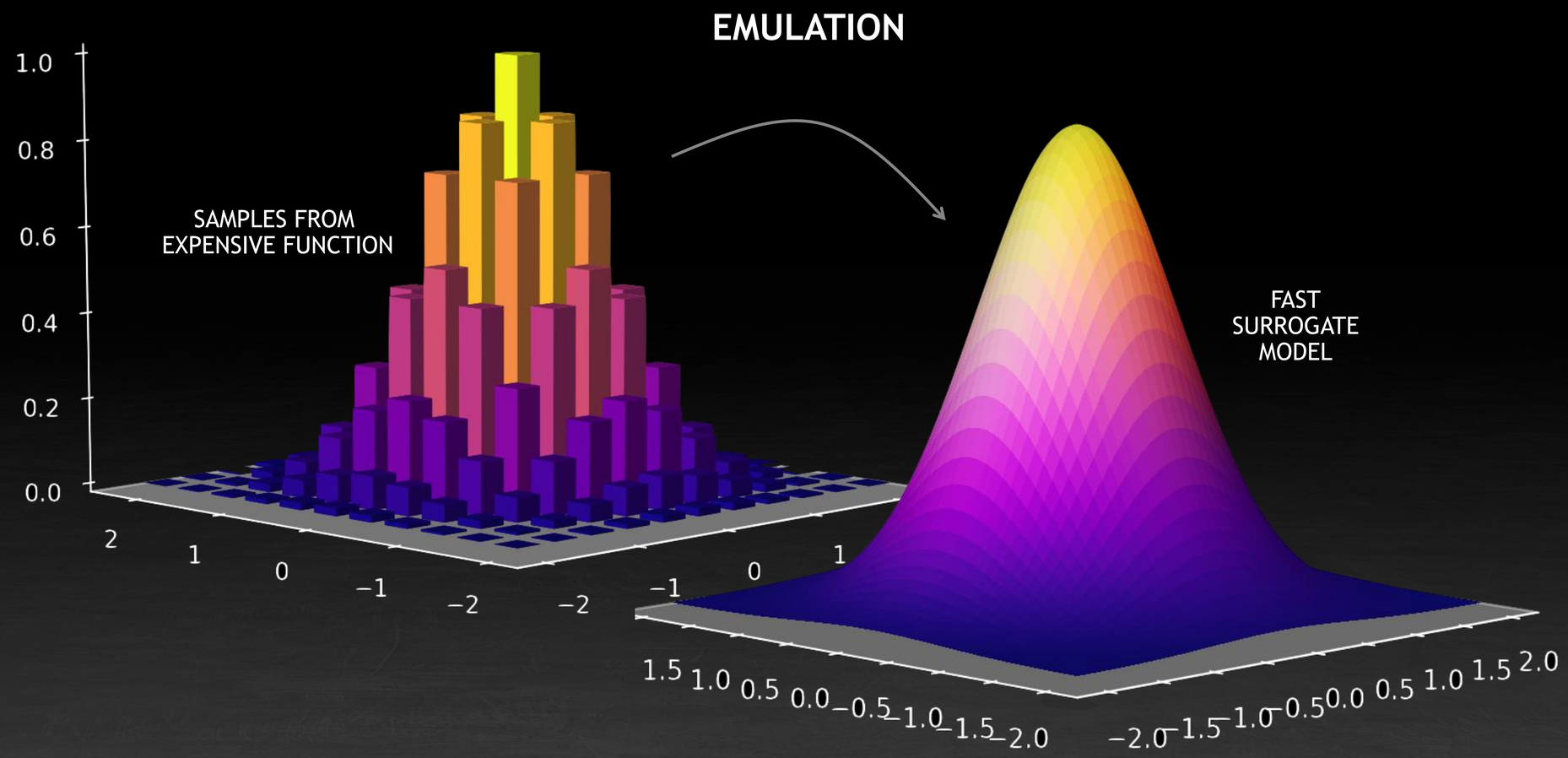

$\mathrm{Log}_{\mathcal{Y}_3}(\mathcal{Y}_2) = \mathcal{V}_2$
$\mathcal{V}_5$
$\mathrm{Log}_{\mathcal{Y}_3}(\mathcal{Y}_4) = \mathcal{V}_4$
$\mathrm{Log}_{\mathcal{Y}_3}(\mathcal{Y}_1) = \mathcal{V}_1$
$T_{\mathcal{Y}_3}\mathcal{M}$
$\mathcal{Y}_3$
$\mathcal{Y}_1$
$\mathcal{Y}_2 = \mathrm{Exp}_{\mathcal{Y}_3}(\mathcal{V}_5)$
$\mathcal{Y}_4$
$\mathcal{M}$

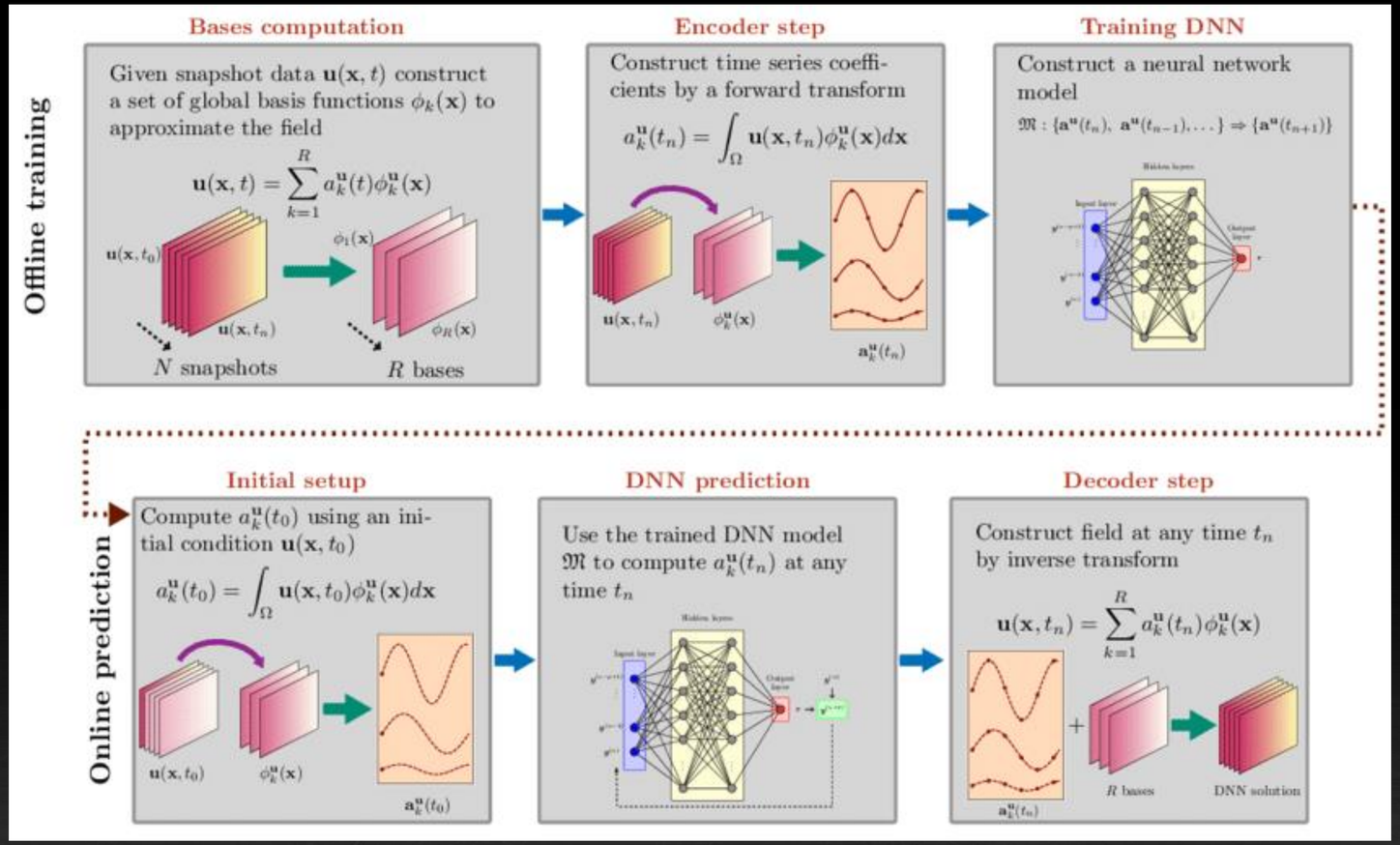

Reduced Space
Manifold
$sE - A$
B
C
D
MOR
$sE_r - A_r$
B_r
C_r
D_r


Reduced Order Methods for Modeling and Computational Reduction
Reduced-Order Modeling (ROM) for Simulation and Optimization
Model Order Reduction
Machine Learning for Model Order Reduction
MODEL ORDER REDUCTION
Reduced-Order Modeling for Flow Control
Machine Learning, Low-Rank Approximations and Reduced Order Modeling in Computational Mechanics
MODEL ORDER REDUCTION
MODEL ORDER REDUCTION TECHNIQUES

ACCELERATED PHYSICS

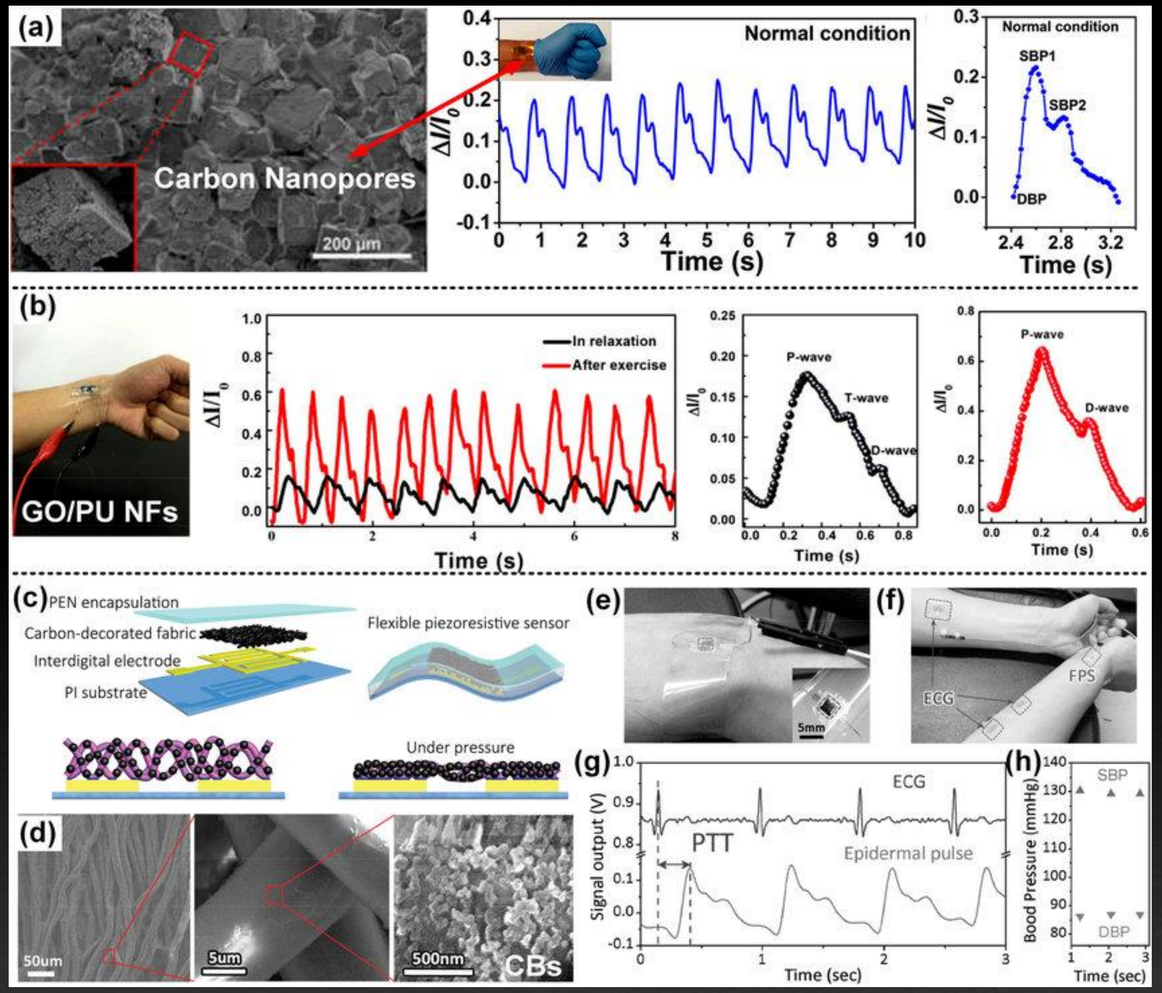
Using surrogate models to speed up existing code

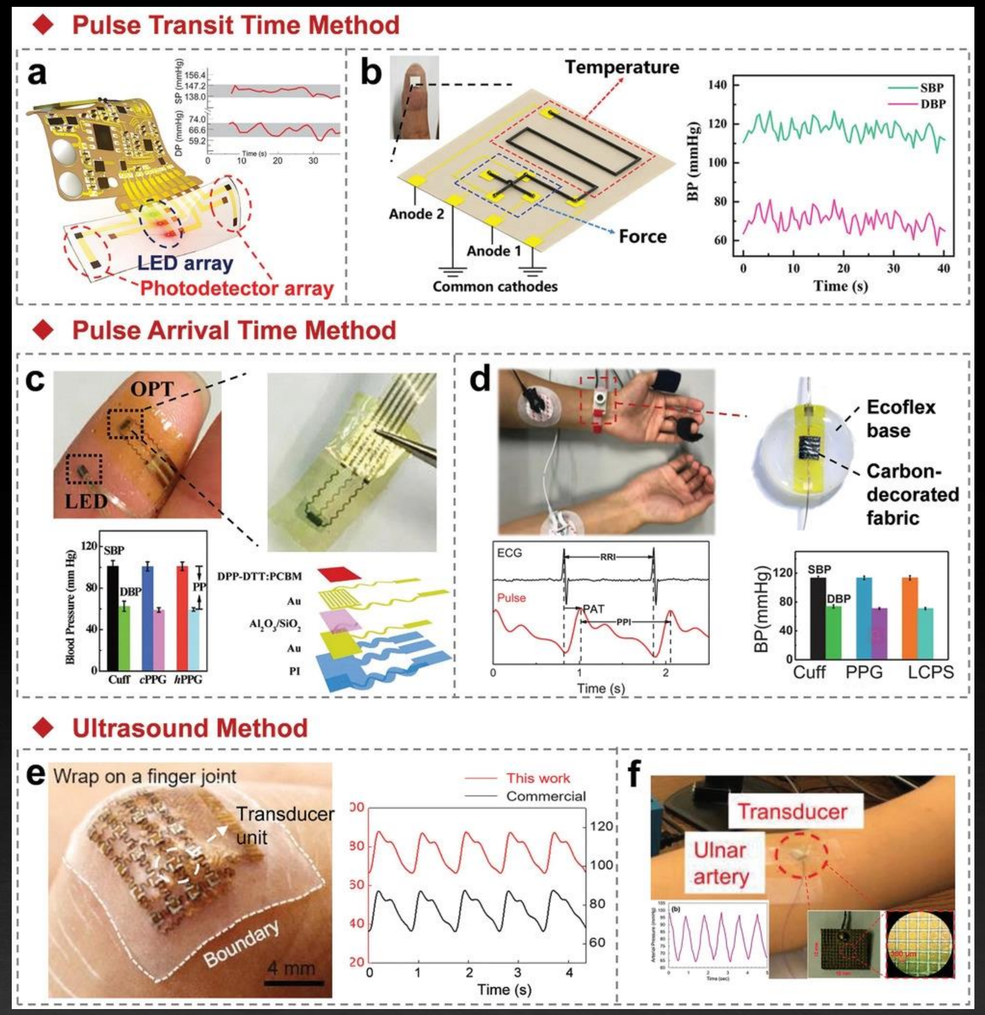


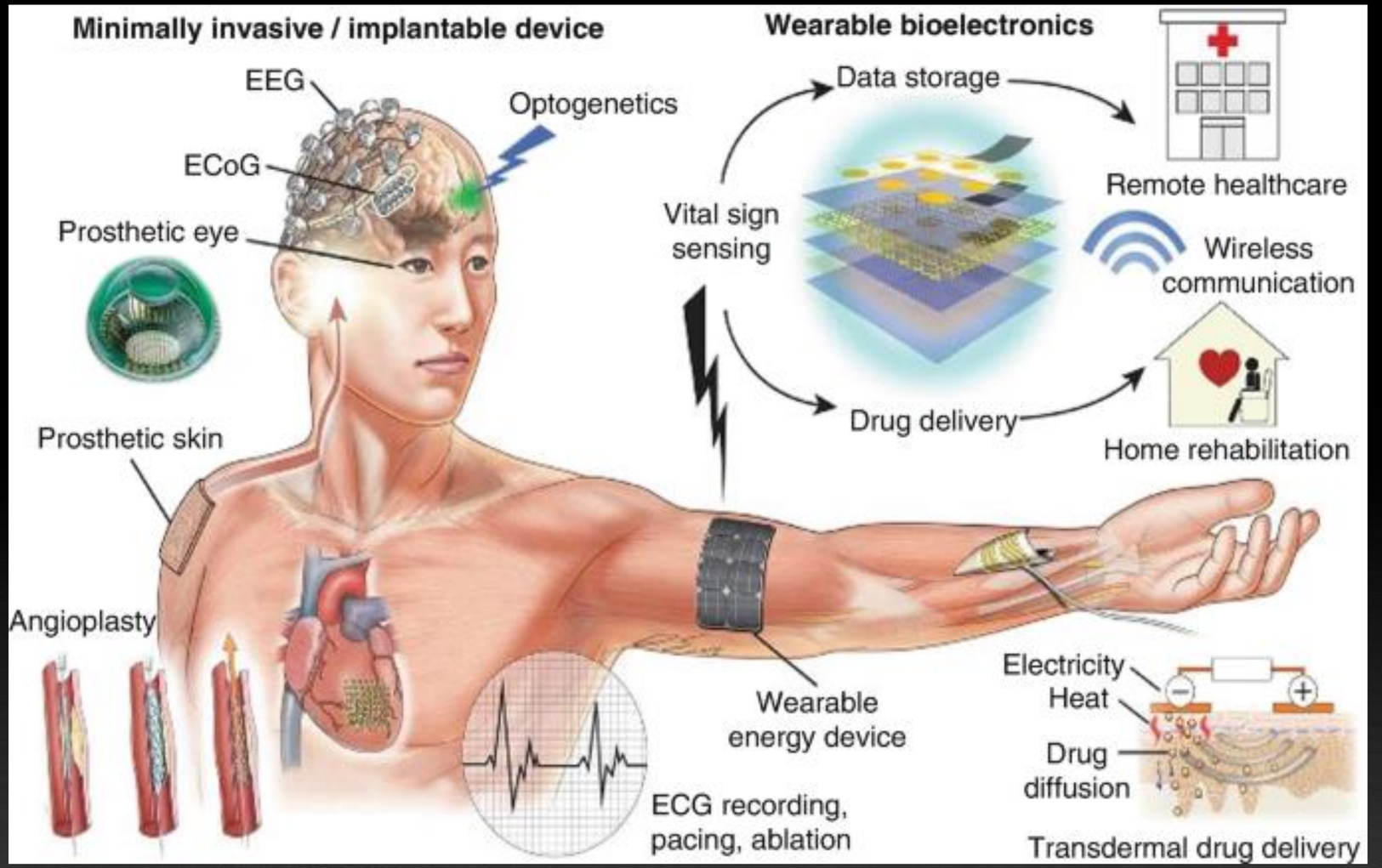

Offline training
Bases computation
Given snapshot data $\mathbf{u}(\mathbf{x},t)$ construct a set of global basis functions $\phi_k(\mathbf{x})$ to approximate the field
$\mathbf{u}(\mathbf{x},t)=\sum_{k=1}^{R} a_k^{\mathbf{u}}(t)\phi_k^{\mathbf{u}}(\mathbf{x})$
$\mathbf{u}(\mathbf{x},t_0)$
$\mathbf{u}(\mathbf{x},t_n)$
$\phi_1(\mathbf{x})$
$\phi_R(\mathbf{x})$
N snapshots
R bases
Encoder step
Construct time series coefficients by a forward transform
$a_k^{\mathbf{u}}(t_n)=\int_\Omega \mathbf{u}(\mathbf{x},t_n)\phi_k^{\mathbf{u}}(\mathbf{x})d\mathbf{x}$
$\mathbf{u}(\mathbf{x},t_n)$
$\phi_k^{\mathbf{u}}(\mathbf{x})$
$\mathbf{a}_k^{\mathbf{u}}(t_n)$
Training DNN
Construct a neural network model
$\mathfrak{M}:\{\mathbf{a}^{\mathbf{u}}(t_n),\ \mathbf{a}^{\mathbf{u}}(t_{n-1}),\dots\}\Rightarrow\{\mathbf{a}^{\mathbf{u}}(t_{n+1})\}$
Online prediction
Initial setup
Compute $a_k^{\mathbf{u}}(t_0)$ using an initial condition $\mathbf{u}(\mathbf{x},t_0)$
$a_k^{\mathbf{u}}(t_0)=\int_\Omega \mathbf{u}(\mathbf{x},t_0)\phi_k^{\mathbf{u}}(\mathbf{x})d\mathbf{x}$
$\mathbf{u}(\mathbf{x},t_0)$
$\phi_k^{\mathbf{u}}(\mathbf{x})$
$\mathbf{a}_k^{\mathbf{u}}(t_0)$
DNN prediction
Use the trained DNN model $\mathfrak{M}$ to compute $a_k^{\mathbf{u}}(t_n)$ at any time t_n
Decoder step
Construct field at any time t_n by inverse transform
$\mathbf{u}(\mathbf{x},t_n)=\sum_{k=1}^{R} a_k^{\mathbf{u}}(t_n)\phi_k^{\mathbf{u}}(\mathbf{x})$
$\mathbf{a}_k^{\mathbf{u}}(t_n)$
R bases
DNN solution


Meisterbüro

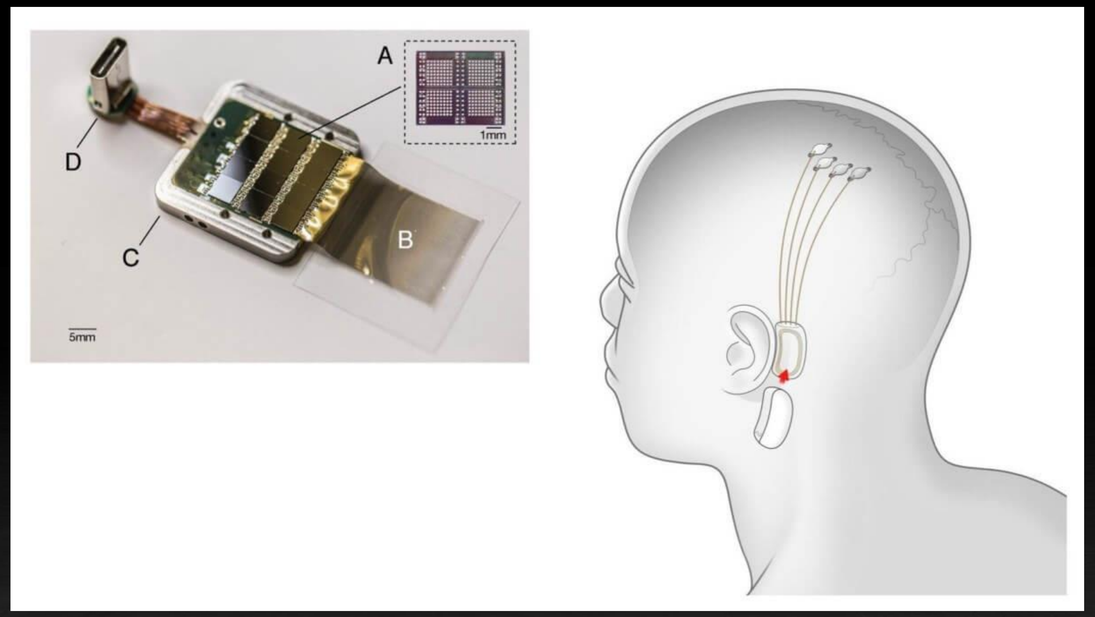

DIGITAL TWIN OF THE HUMAN


(a)
Carbon Nanopores
200 μm
Normal condition
SBP1
SBP2
DBP
Time (s)
(b)
GO/PU NFs
In relaxation
After exercise
P-wave
T-wave
D-wave
Time (s)
(c)
PEN encapsulation
Carbon-decorated fabric
Interdigital electrode
PI substrate
Flexible piezoresistive sensor
Under pressure
(d)
50um
5um
500nm
CBs
(e)
5mm
(f)
ECG
FPS
(g)
ECG
PTT
Epidermal pulse
Signal output (V)
Time (sec)
(h)
SBP
DBP
Bood Pressure (mmHg)
Time (sec)

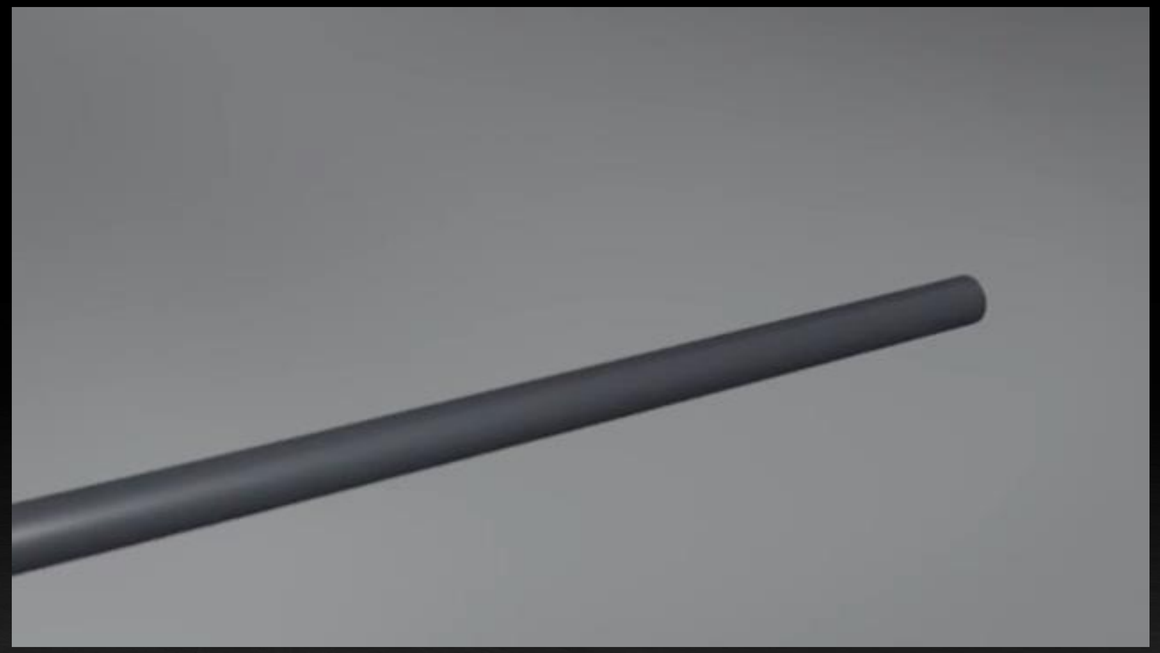

Pulse Transit Time Method
a
LED array
Photodetector array
b
Temperature
Anode 2
Anode 1
Force
Common cathodes
SBP
DBP
BP (mmHg)
Time (s)
Pulse Arrival Time Method
c
OPT
LED
Blood Pressure (mm Hg)
Cuff
cPPG
hPPG
DPP-DTT:PCBM
Au
Al_2O_3/SiO_2
Au
PI
d
Ecoflex base
Carbon-decorated fabric
ECG
RRI
Pulse
PAT
PPI
Time (s)
BP(mmHg)
Cuff
PPG
LCPS
Ultrasound Method
e
Wrap on a finger joint
Transducer unit
Boundary
4 mm
This work
Commercial
Time (s)
f
Transducer
Ulnar artery


Minimally invasive / implantable device
Wearable bioelectronics
EEG
Optogenetics
ECoG
Prosthetic eye
Prosthetic skin
Angioplasty
Data storage
Vital sign sensing
Drug delivery
Remote healthcare
Wireless communication
Home rehabilitation
Wearable energy device
ECG recording, pacing, ablation
Electricity
Heat
Drug diffusion
Transdermal drug delivery

BRAIN COMPUTER INTERFACE

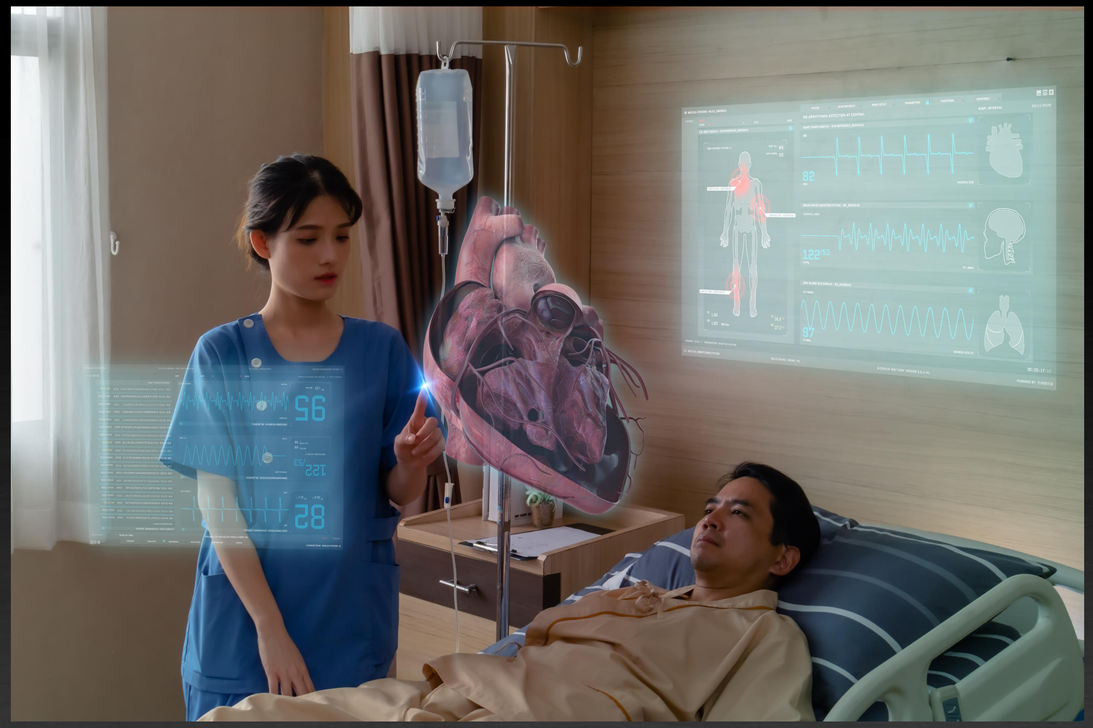


NEURALINK



SYNCHRON

HUMAN DIGITAL TWINS





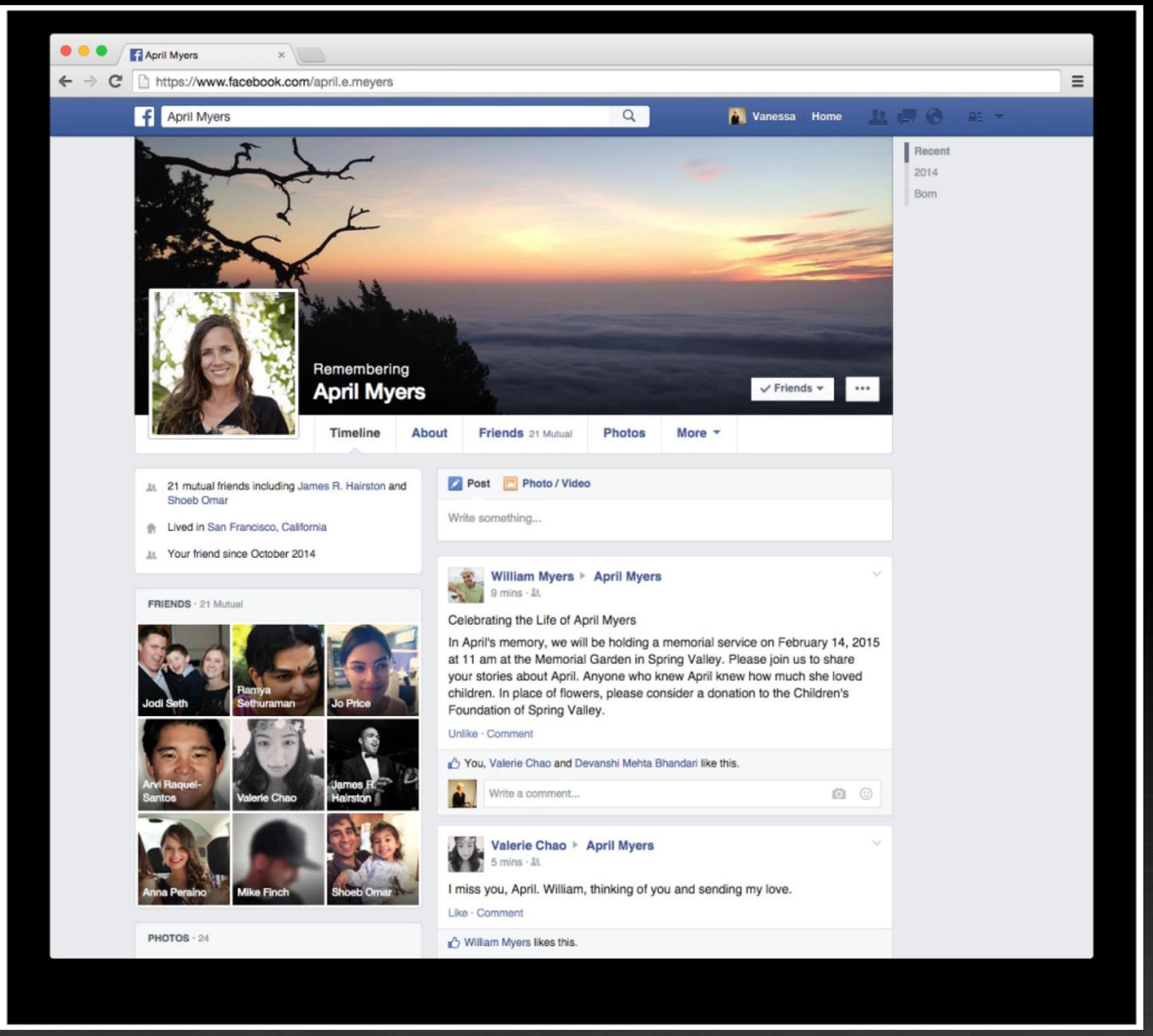


Simulating the Deceased



Science Fiction precedes Science Facts




April Myers
https://www.facebook.com/april.e.meyers
April Myers
Vanessa
Home
Recent
2014
Born
Remembering
April Myers
Friends
Timeline
About
Friends 21 Mutual
Photos
More
21 mutual friends including James R. Hairston and Shoeb Omar
Lived in San Francisco, California
Your friend since October 2014
FRIENDS · 21 Mutual
Jodi Seth
Raniya Sethuraman
Jo Price
Arvi Raquel-Santos
Valerie Chao
James R. Hairston
Anna Peraino
Mike Finch
Shoeb Omar
PHOTOS · 24
Post
Photo / Video
Write something...
William Myers ▸ April Myers
9 mins ·
Celebrating the Life of April Myers
In April's memory, we will be holding a memorial service on February 14, 2015 at 11 am at the Memorial Garden in Spring Valley. Please join us to share your stories about April. Anyone who knew April knew how much she loved children. In place of flowers, please consider a donation to the Children's Foundation of Spring Valley.
Unlike · Comment
You, Valerie Chao and Devanshi Mehta Bhandari like this.
Write a comment...
Valerie Chao ▸ April Myers
5 mins ·
I miss you, April. William, thinking of you and sending my love.
Like · Comment
William Myers likes this.




엄마 나연이 너무 보고 싶었어

Grieving man uses AI site to ‘chat’ with dead girlfriend

By Lee Brown

July 26, 2021 | 4:19pm | Updated



Joshua Barbeau used GPT-3, an artificial intelligence software, to simulate a chat with his deceased girlfriend Jessica Pereira.

Facebook

Record Review

Francesca Motti



Tell me a story about your childhood.
Something I don't know...

Home Privacy policy Feedback



Tech

Microsoft patent shows plans to revive dead loved ones as chatbots

The patent also mentions using 2D or 3D models of specific people

Adam Smith • Wednesday 20 January 2021 12:15 • Comments



US010853717B2

(12) **United States Patent**
Abramson et al.

(10) **Patent No.: US 10,853,717 B2**
(45) **Date of Patent: Dec. 1, 2020**

(54) **CREATING A CONVERSATIONAL CHAT BOT OF A SPECIFIC PERSON**

(71) Applicant: **Microsoft Technology Licensing, LLC**, Redmond, WA (US)

(72) Inventors: **Dustin I Abramson**, Bellevue, WA (US); **Joseph Johnson, Jr.**, Seattle, WA (US)

(73) Assignee: **Microsoft Technology Licensing, LLC**, Redmond, WA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 873 days.

(21) Appl. No.: **15/484,470**

(22) Filed: **Apr. 11, 2017**

(65) **Prior Publication Data**

US 2018/0293483 A1 Oct. 11, 2018

(51) **Int. Cl.**
G06N 3/00 (2006.01)
H04L 12/58 (2006.01)
G06N 20/00 (2019.01)

(52) **U.S. Cl.**
CPC ***G06N 3/006*** (2013.01); ***G06N 20/00*** (2019.01); ***H04L 51/02*** (2013.01); *H04L 51/04* (2013.01); *H04L 51/32* (2013.01)

(58) **Field of Classification Search**
CPC G06F 16/3329
USPC 706/15, 45
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

8,433,344 B1 * 4/2013 Virga G09B 29/106 455/457
8,719,200 B2 5/2014 Beilby et al.
8,819,549 B2 8/2014 Nageswaram et al.
9,514,748 B2 12/2016 Reddy et al.
2002/0010584 A1 1/2002 Schultz et al.
2009/0254417 A1 * 10/2009 Beilby G06N 3/004 706/45
2013/0257877 A1 10/2013 Davis
(Continued)

FOREIGN PATENT DOCUMENTS

WO 2003073417 A2 9/2003

OTHER PUBLICATIONS

Wang, et al., "High Quality Lip-Sync Animation for 3d Photo-Realistic Talking Head", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 25, 2012, pp. 4529-4532.
(Continued)

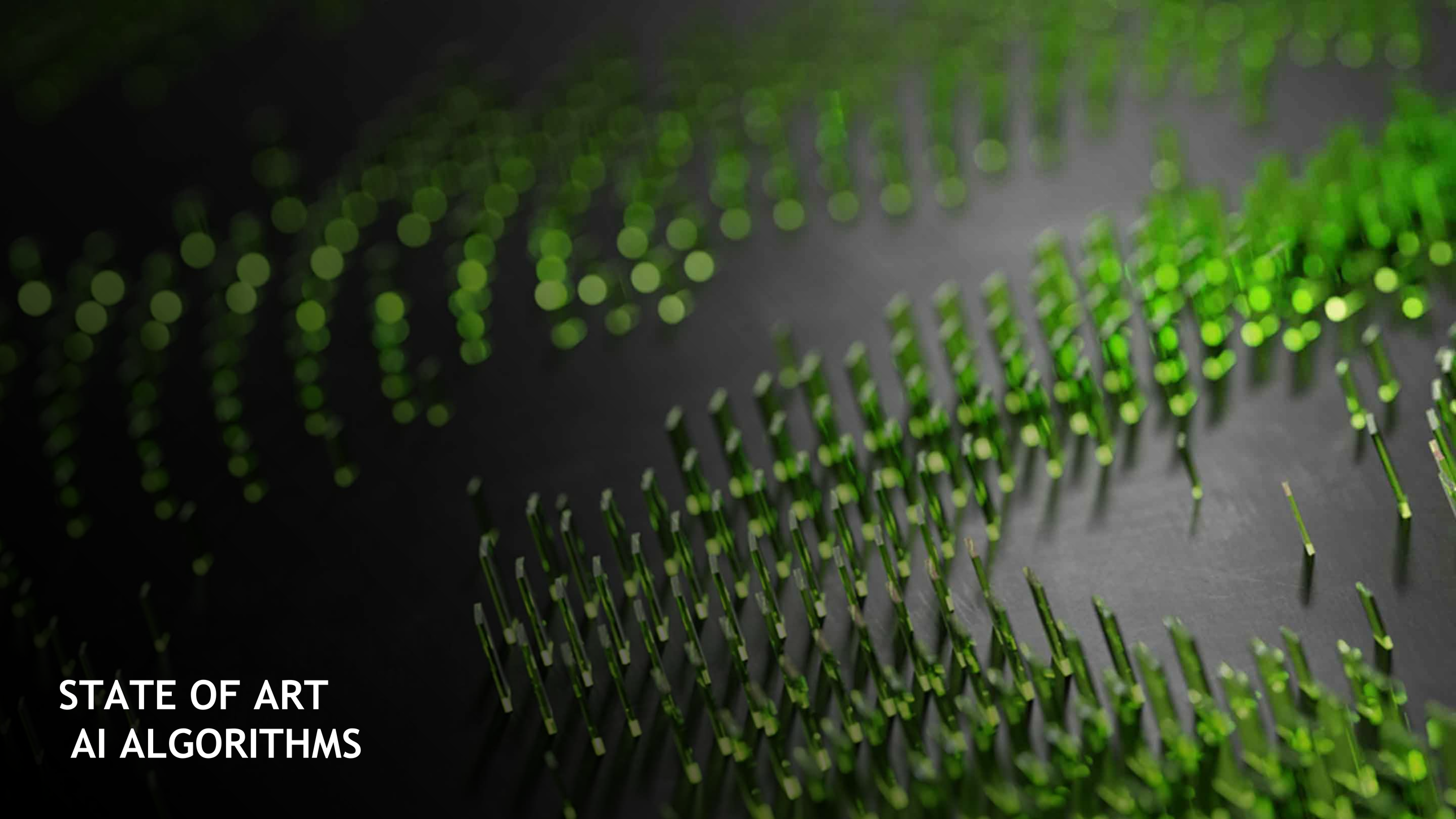
Primary Examiner — David R Vincent
(74) *Attorney, Agent, or Firm* — Merchant & Gould P.C.

(57) **ABSTRACT**

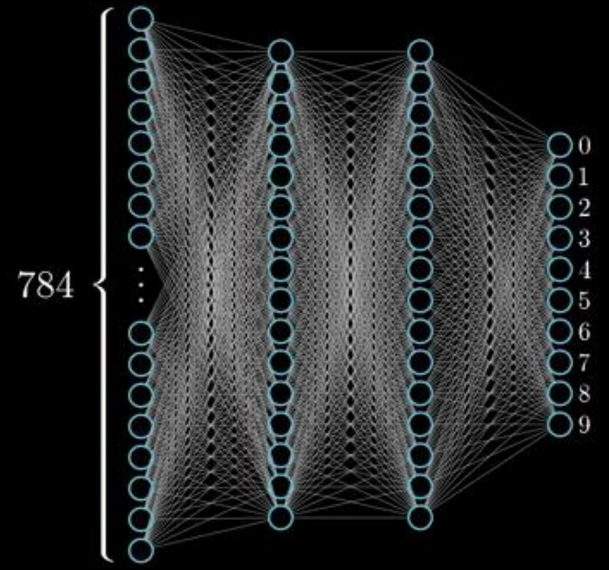
Examples of the present disclosure describe systems and methods of creating a conversational chat bot of a specific person. In aspects, social data (e.g., images, voice data, social media posts, electronic messages, written letters, etc.) about the specific person may be accessed. The social data may be used to create or modify a special index in the theme of the specific person's personality. The special index may be used to train a chat bot to converse in the personality of the specific person. During such conversations, one or more conversational data stores and/or APIs may be used to reply to user dialogue and/or questions for which the social data does not provide data. In some aspects, a 2D or 3D model of a specific person may be generated using images, depth information, and/or video data associated with the specific person.

20 Claims, 8 Drawing Sheets




STATE OF ART
AI ALGORITHMS



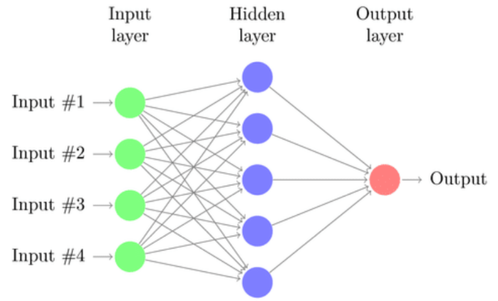


Theorem (*Cybenko, 1989*)

Let σ be any continuous sigmoidal function. Then, the finite sums of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j \cdot x + \theta_j)$$

are dense in $C(I_d)$.



Theorem 4.2. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous with respect to the Lebesgue measure with Radon–Nikodym density $\rho(x)$. density $\rho(x)$. Let $\nu = \sum_{i=1}^{n} \nu_i \delta_{y_i}$. Let $\psi = (\psi_1, \cdots, \psi_n)$ be an maximizer of $\mathcal{F}$ defined in (4.7). Denote by $\{P_j\}_{j=1}^{n}$ the power diagrams associated to $\{y_j\}_{j=1}^{n}$ and ψ. Then the optimal transport plan T solving the semi-discrete Monge's problem (4.4) is given by $T(x) = \nabla\bar{\varphi}(x)$, where $\bar{\varphi}(x) = \max_j\{x \cdot y_j + m_j\}$ for some $m_j \in \mathbb{R}$. Specifically, $T(x) = y_j$ if $x \in P_j(\psi)$.*

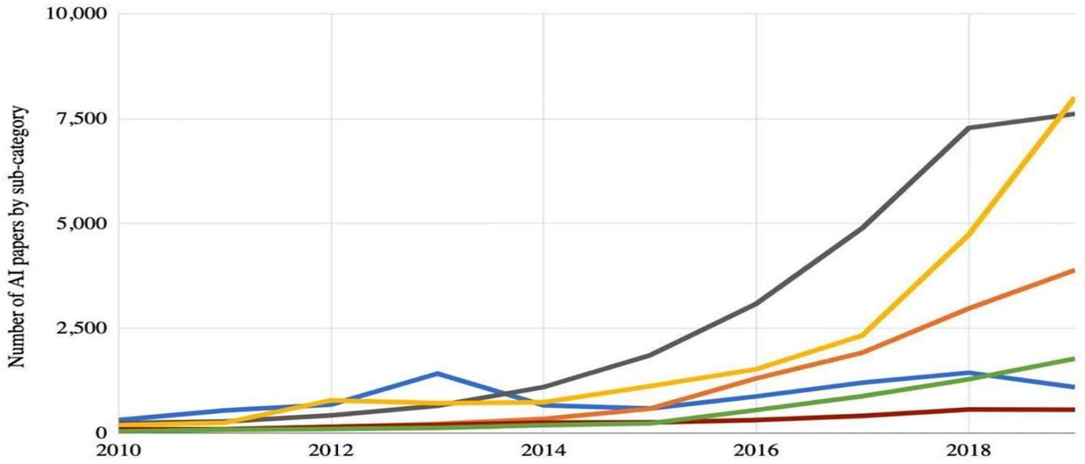
A Universal Approximation Theorem of Deep Neural Networks for Expressing Probability Distributions

Yulong Lu
Department of Mathematics and Statistics
University of Massachusetts Amherst
Amherst, MA 01003
lu@math.umass.edu

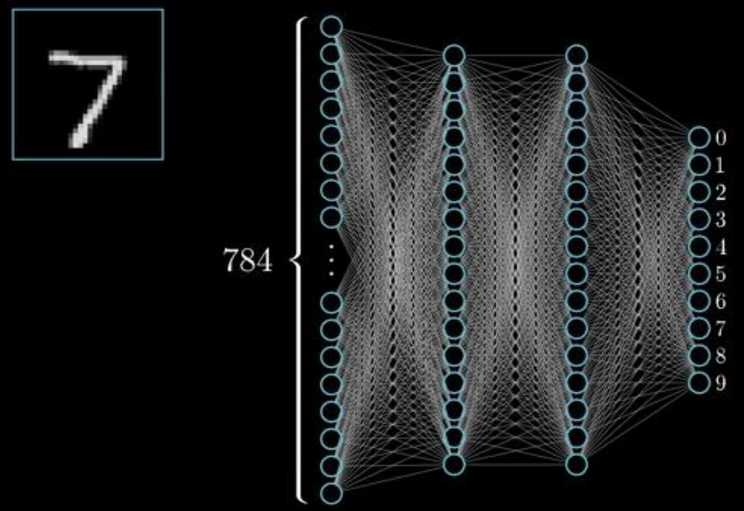
Jianfeng Lu
Mathematics Department
Duke University
Durham, NC 27708
jianfeng@math.duke.edu

Number of AI papers on arXiv, 2010-2019

Source: arXiv, 2019.



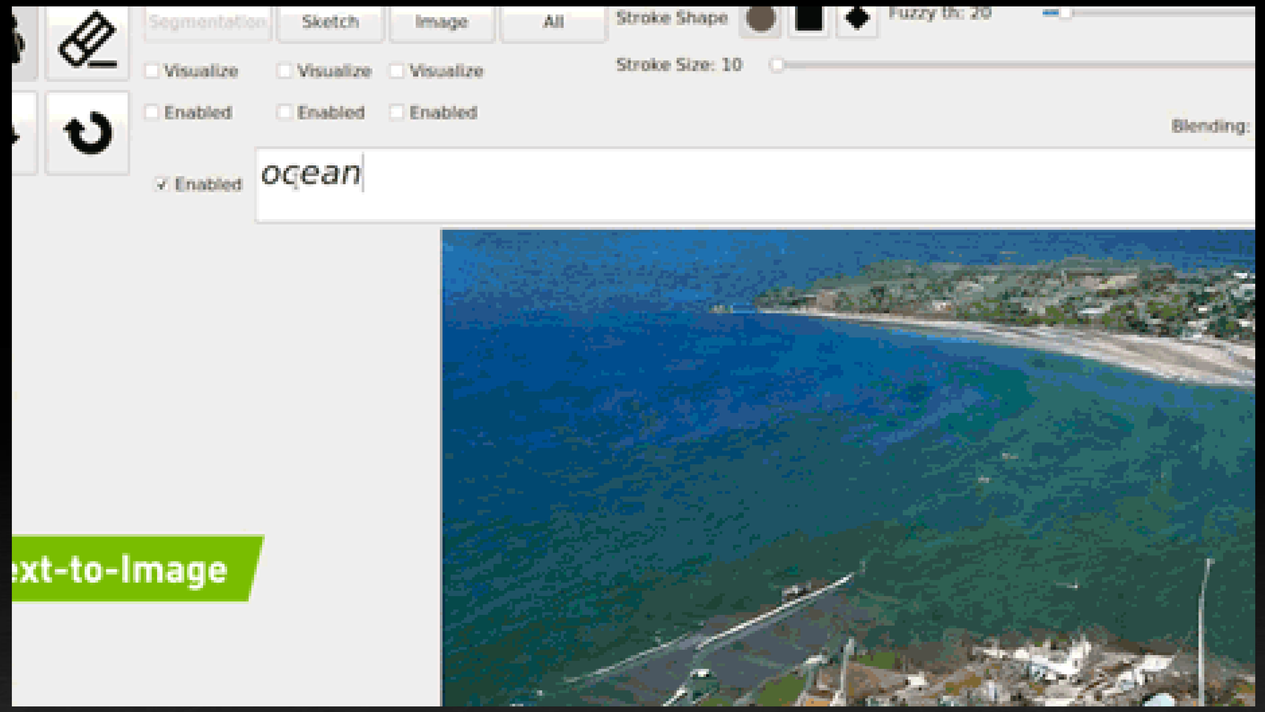
Streamline Machine Learning
www.Spell.run





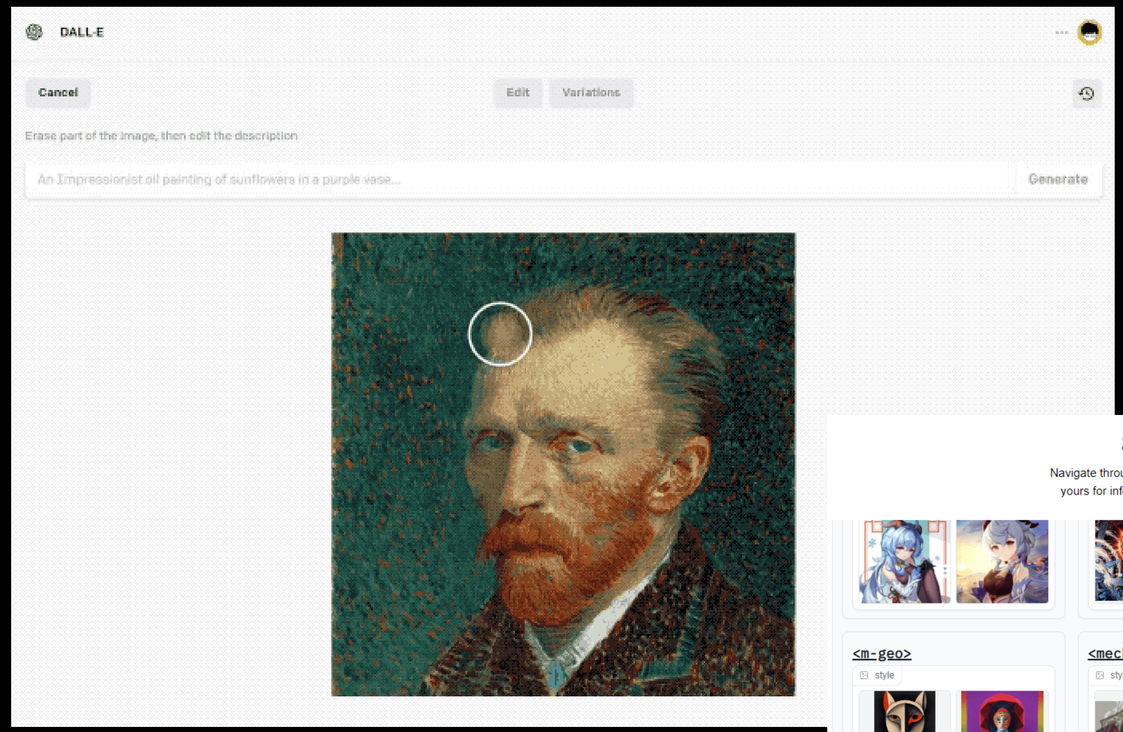


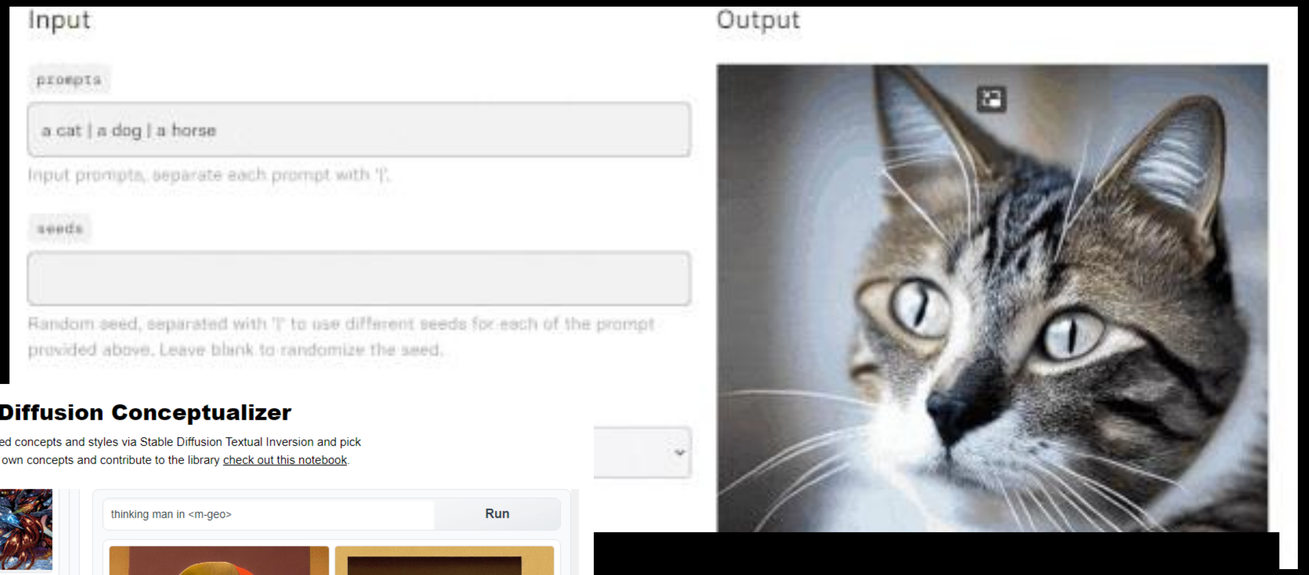


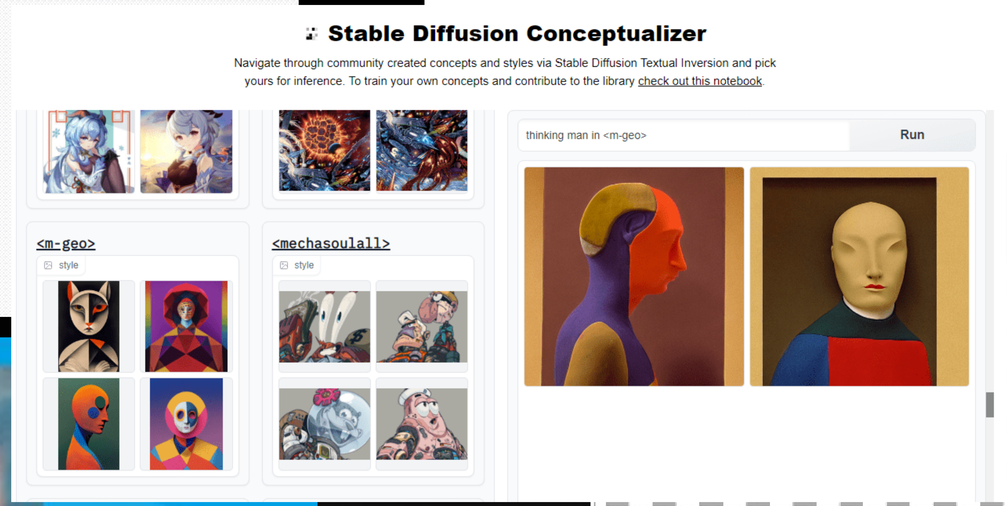


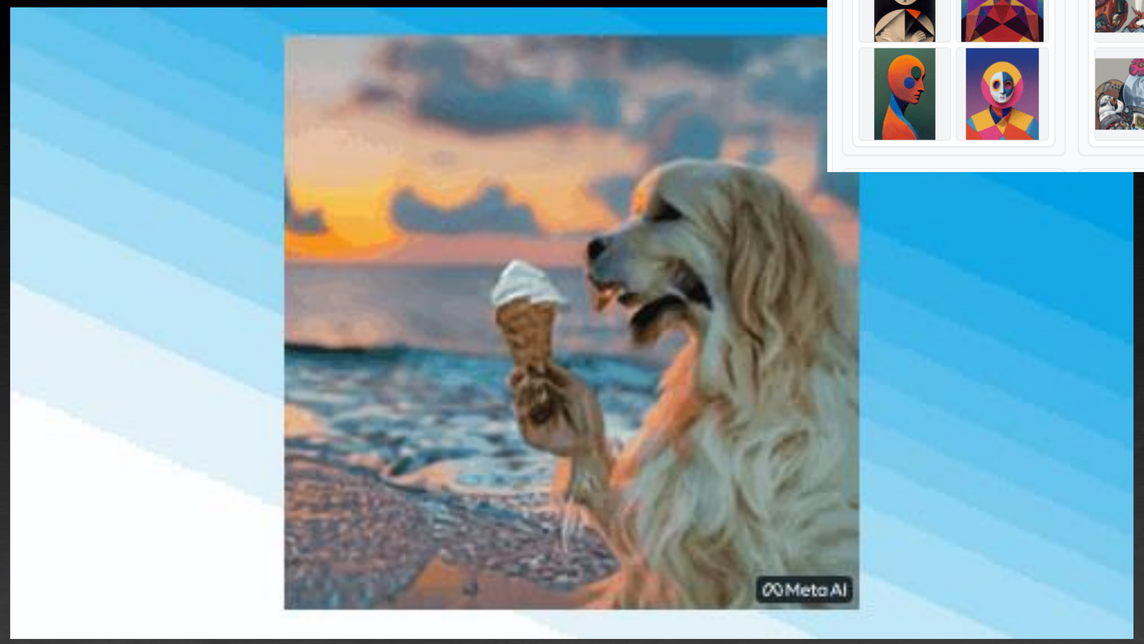


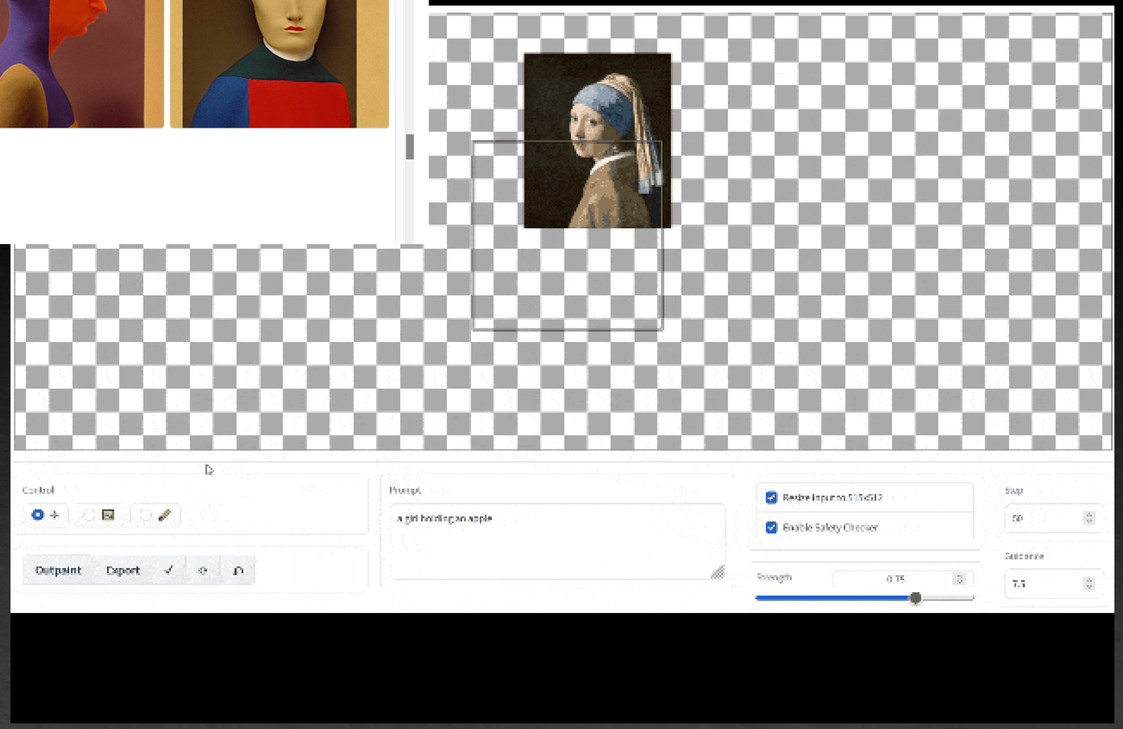
GAUGAN 2AND GAUGAN360


DALL-E
Cancel
Edit
Variations
Erase part of the image, then edit the description
An Impressionist oil painting of sunflowers in a purple vase...
Generate


Input
prompts
a cat | a dog | a horse
Input prompts, separate each prompt with '|'.
seeds
Random seed, separated with '|' to use different seeds for each of the prompt provided above. Leave blank to randomize the seed.
Output


Stable Diffusion Conceptualizer
Navigate through community created concepts and styles via Stable Diffusion Textual Inversion and pick yours for inference. To train your own concepts and contribute to the library check out this notebook.
<m-geo>
style
<mechasoulall>
style
thinking man in <m-geo>
Run

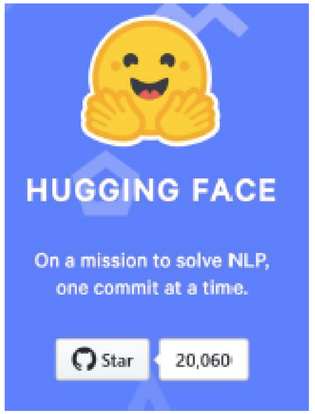

Meta AI


Control
Prompt
a girl holding an apple
Enable Safety Checker
Outpaint
Export
Strength
Guidance



Transformer-Based Language Models

- BERT (Google)
- XLNet (Google/CMU)
- RoBERTa (Facebook)
- DistilBERT (HuggingFace)
- CTRL (Salesforce)
- GPT-2 (OpenAI)
- ALBERT (Google)
- Megatron (NVIDIA)



sebastianruder / NLP-progress

NLP-progress

Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks.

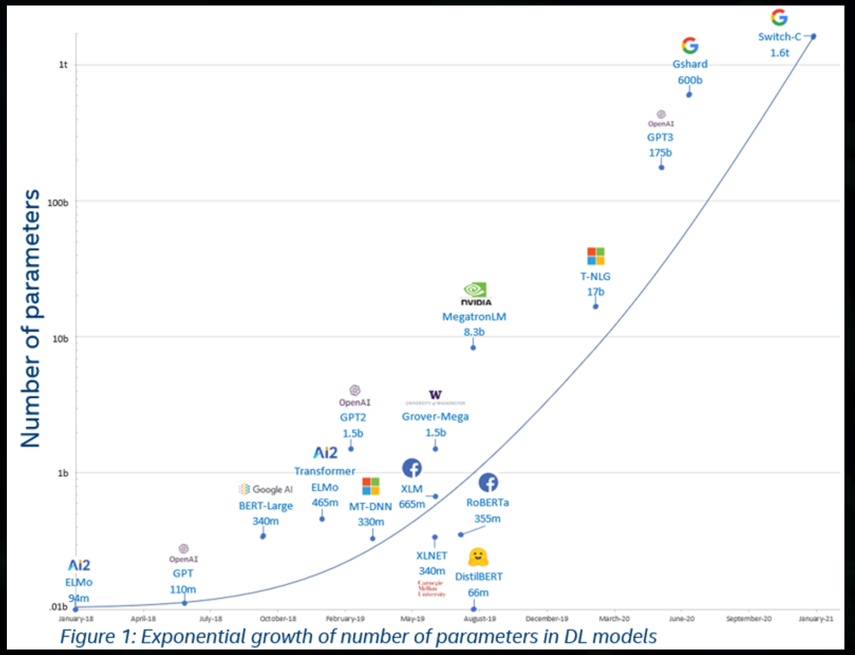
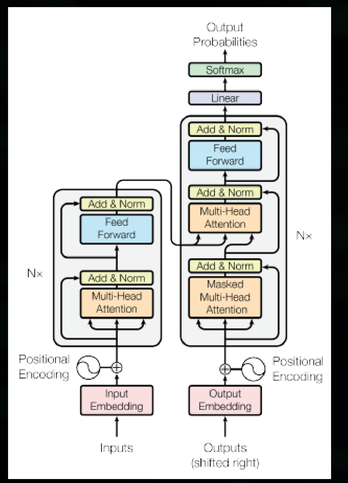


Figure 1: Exponential growth of number of parameters in DL models



$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}}$$

equivalently:

$$PP(W) = 2^{-l}$$

where $l = \frac{1}{N} \log P(w_1 w_2 ... w_N)$

$$2^{-l} \text{ where } l = \frac{1}{M} \sum_{i=1}^{m} \log p(s_i)$$

CONVERSATIONAL AI IS TRANFORMING INDUSTRIES



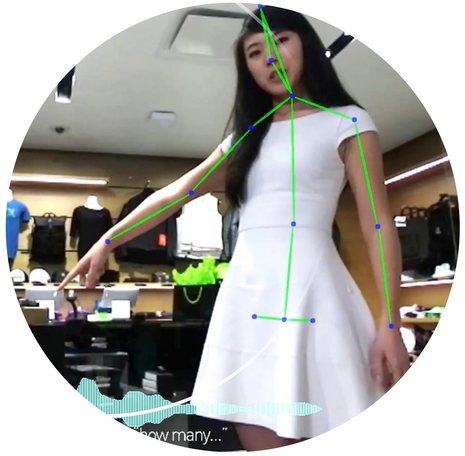
VIDEOCONFERENCE
CC, TRANSLATION, TRANSCRIPTION
200M Meetings per Day



CALL CENTER
500M Calls per Day



SMART SPEAKERS
150M Sold per Year



RETAIL ASSISTANTS
12M Retail Stores



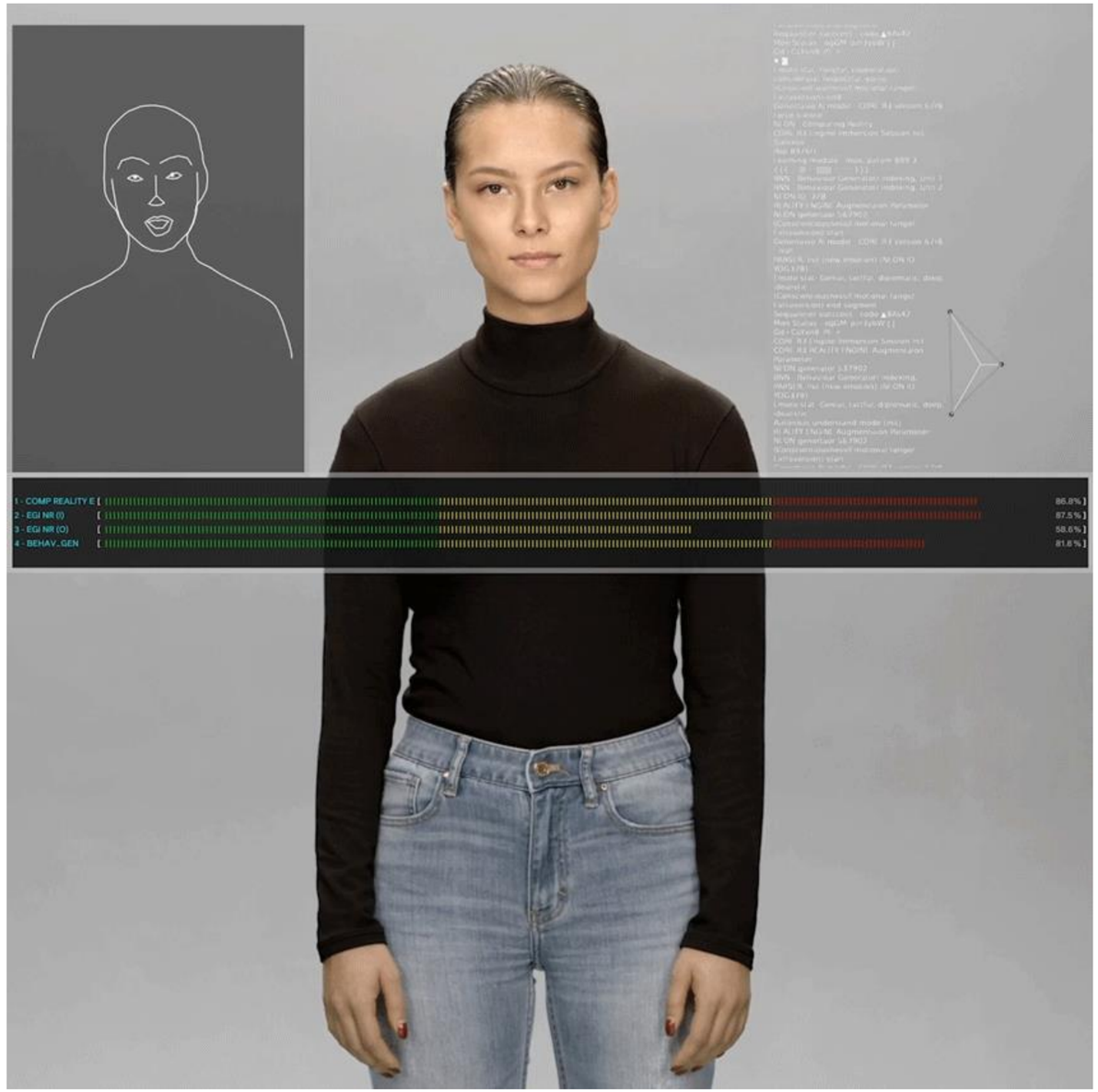
IN-CAR ASSISTANTS
75M New Cars per Year

Samsung's new Neon project is finally unveiled: It's a humanoid AI chatbot









Artificial intelligence / Machine learning

OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

by **Will Douglas Heaven** July 20, 2020

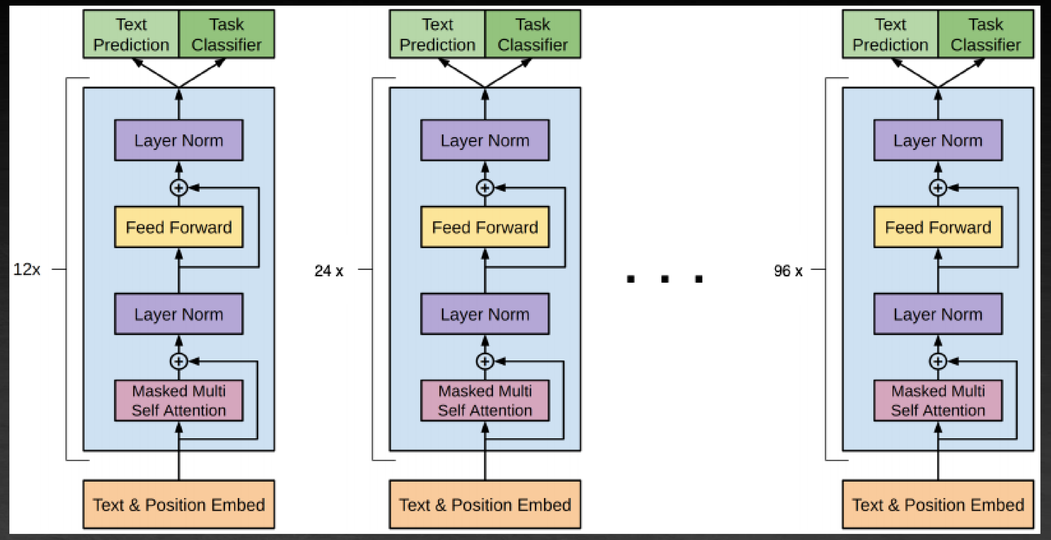
A robot wrote this entire article. Are you scared yet, human?

GPT-3

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below

Tue 8 Sep 2020 09.45 BST



TECH ARTIFICIAL INTELLIGENCE

OpenAI's text-generating system GPT-3 is now spewing out 4.5 billion words a day

Robot-generated writing looks set to be the next big thing

By James Vincent | Mar 29, 2021, 8:24am EDT



Make it be vertically centered; put on the left side of the page.

```
var rocketship =
document.createElement('img');
rocketship.src =
'https://i1.sndcdn.com/artwork
s-j8xjG7zc1wmTeO7b-O6l83w-
t500x500.jpg';
document.body.appendChild(rock
etship);


/* Make it be smallish. */
rocketship.style.width =
'100px';
rocketship.style.height =
'100px';


/* Crop it circularly. */
rocketship.style.borderRadius
= '50%';


/* Make it be vertically
centered; put on the left side
of the page. */
rocketship.style.position =
'absolute';
rocketship.style.left = '0px';
rocketship.style.top = '50%';
rocketship.style.marginTop =
'-50px';
```



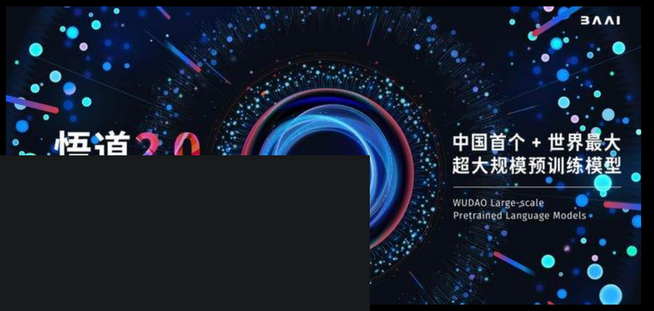
北京智源人工智能研究院
BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

超越图灵测试的AI作诗

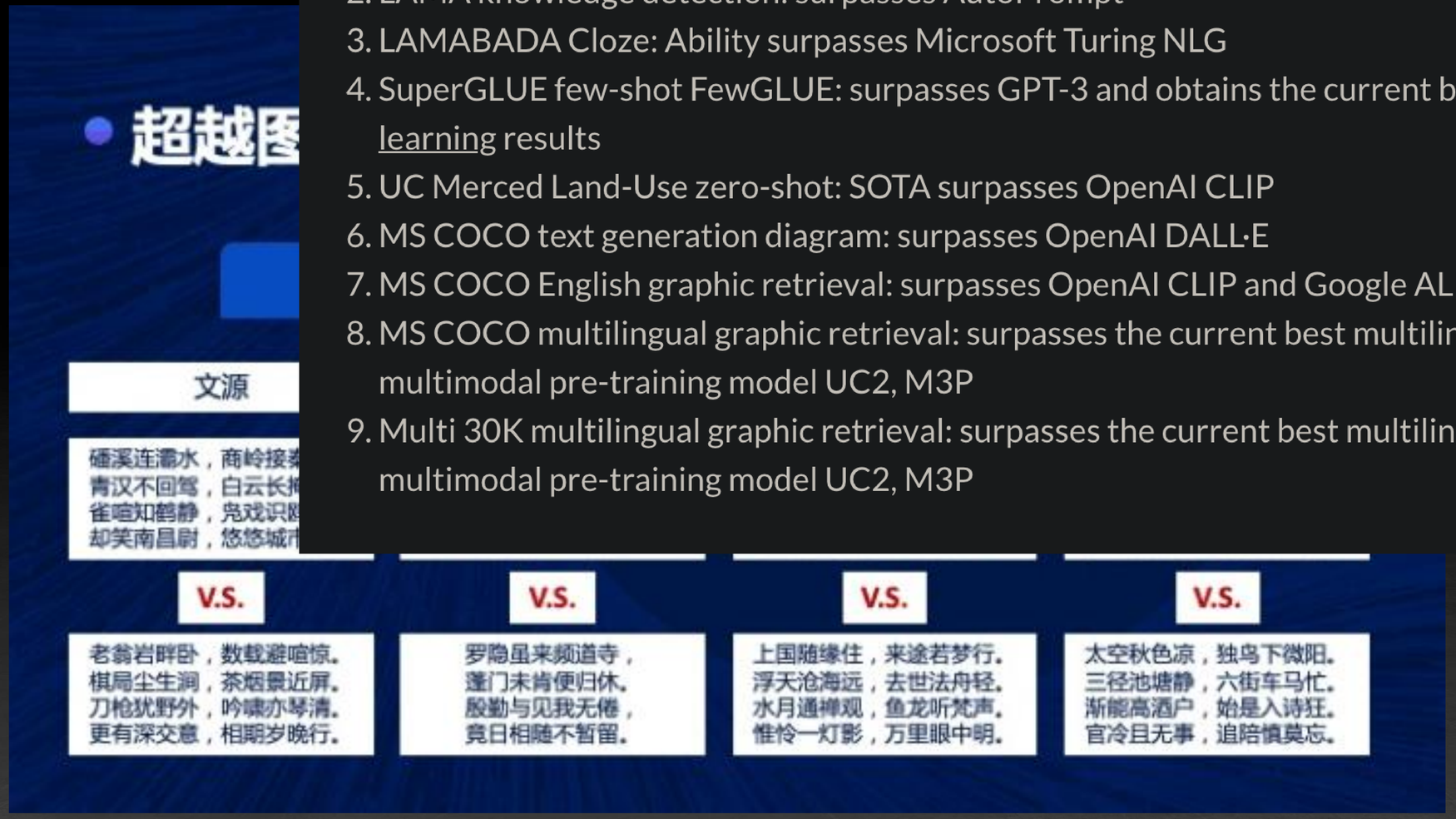
https://turing-poet.aminer.cn/index.html

悟道 V.S. 唐宋诗人

文源	文澜	文溯	文汇
碛溪连灞水，商岭接秦山。 青汉不回驾，白云长掩关。 雀喧知鹤静，鸟戏识鸥闲。 却笑南昌尉，悠悠城市间。	晨趋禁掖暮郊园， 松桂苍苍烟露繁。 明月上时群动息， 雪峰高处正当轩。	初辞云浦寂，再别竺岩难。 各分当时意，空驰海远寒。 依稀真弟子，守至四明滩。 临别殷勤语，重违誓去盘。	近年容尚白，卿士累资名。 投契顾怜旧，照乘拟次行。 常嫌官爵小，更睹道心清。 多谢白公意，作为肺腑情。
V.S.	V.S.	V.S.	V.S.
老翁岩畔卧，数载避喧惊。 棋局尘生润，茶烟景近屏。 刀枪犹野外，吟啸亦琴清。 更有深交意，相期岁晚行。	罗隐虽来频道寺， 蓬门未肯便归休。 殷勤与见我无倦， 竟日相随不暂留。	上国随缘住，来途若梦行。 浮天沧海远，去世法舟轻。 水月通禅观，鱼龙听梵声。 惟怜一灯影，万里眼中明。	太空秋色凉，独鸟下微阳。 三径池塘静，六街车马忙。 渐能高酒户，始是入诗狂。 官冷且无事，追陪慎莫忘。



1. ImageNet zero-shot: SOTA surpasses OpenAI CLIP
2. LAMA knowledge detection: surpasses AutoPrompt
3. LAMABADA Cloze: Ability surpasses Microsoft Turing NLG
4. SuperGLUE few-shot FewGLUE: surpasses GPT-3 and obtains the current best few-shot learning results
5. UC Merced Land-Use zero-shot: SOTA surpasses OpenAI CLIP
6. MS COCO text generation diagram: surpasses OpenAI DALL·E
7. MS COCO English graphic retrieval: surpasses OpenAI CLIP and Google ALIGN
8. MS COCO multilingual graphic retrieval: surpasses the current best multilingual and multimodal pre-training model UC2, M3P
9. Multi 30K multilingual graphic retrieval: surpasses the current best multilingual and multimodal pre-training model UC2, M3P

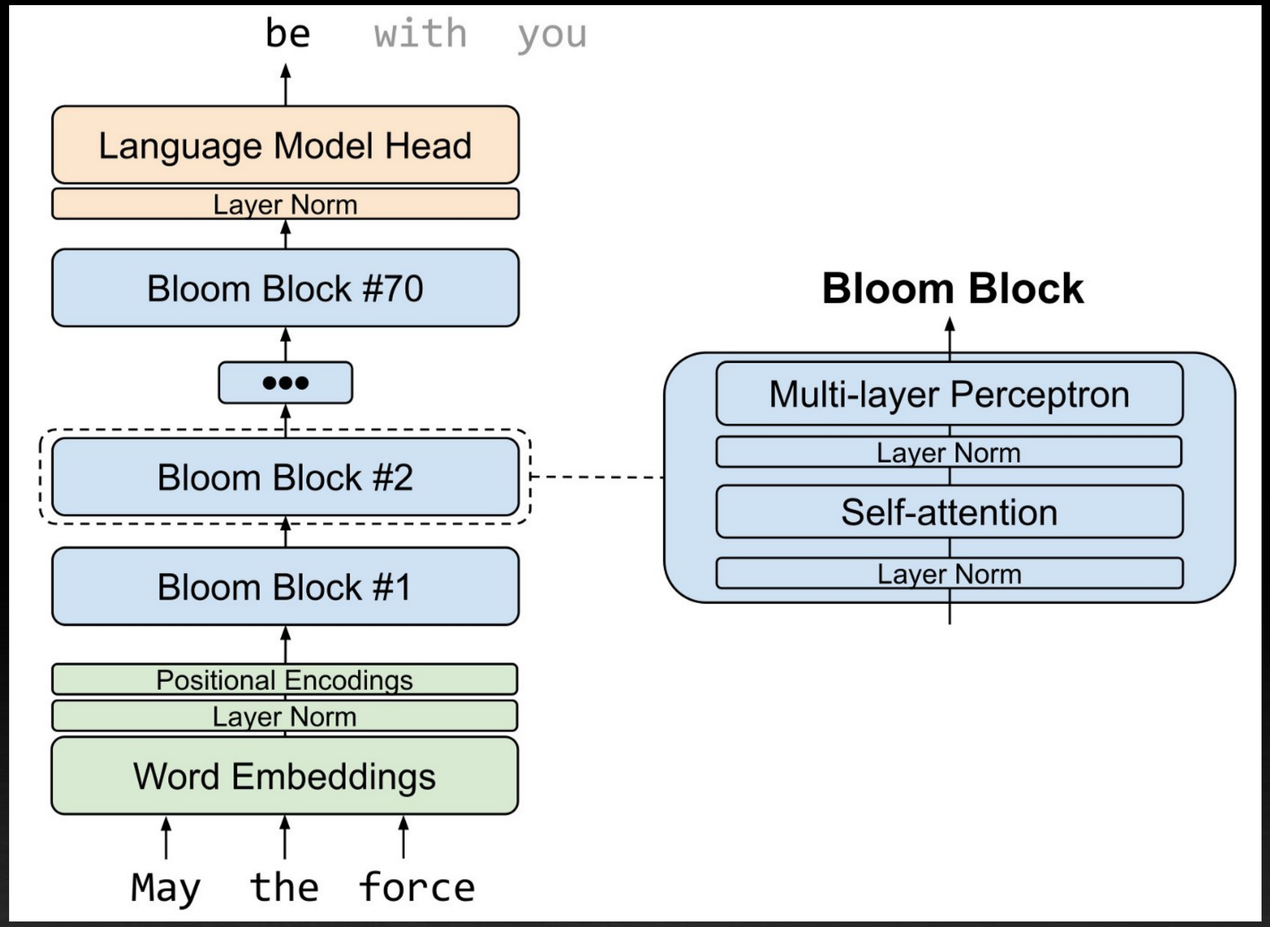


Understand BLOOM, the Largest Open-Access AI, and Run It on Your Local Computer

See BLOOM in action solving math, translation, and coding problems.

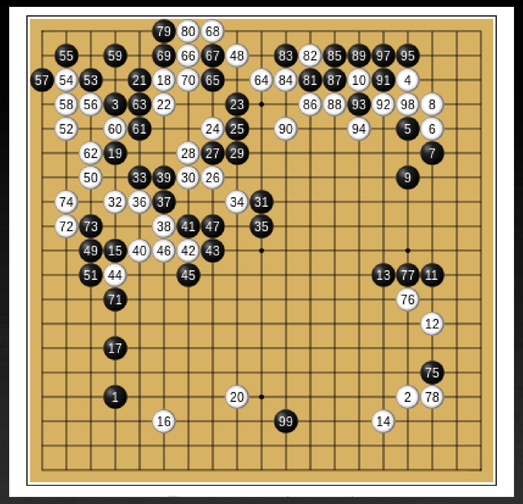


BLOOM is an open-access multilingual language model that contains 176 billion parameters and was trained for 3.5 months on 384 A100–80GB GPUs. A BLOOM checkpoint takes 330 GB of disk space, so it seems unfeasible to run this model on a desktop computer. However, you just need enough disk space, at least 16GB of RAM, and some patience (you don't even need a GPU), to run this model on your computer.



AlphaGo

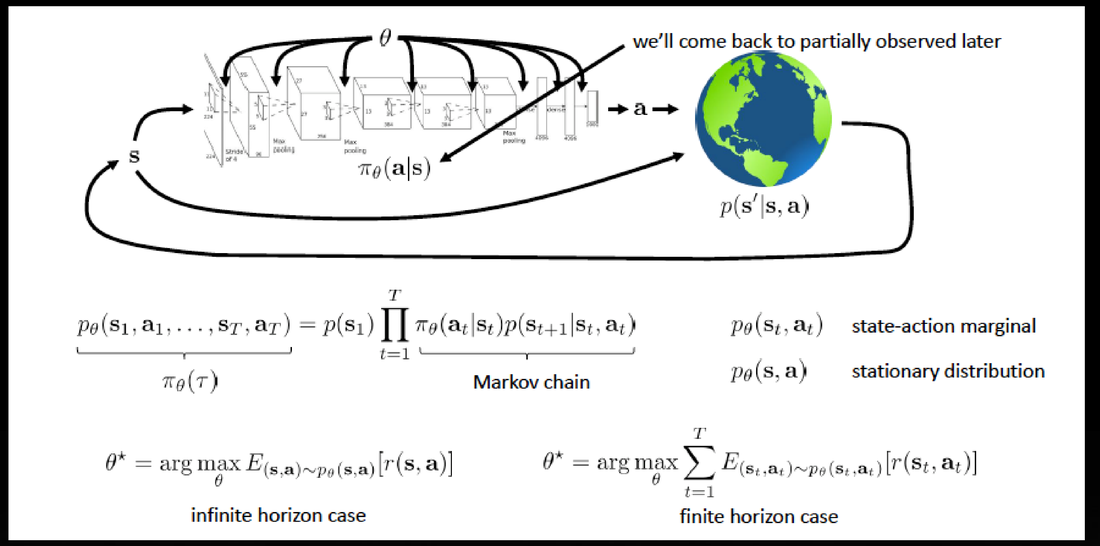






AI: ALPHAGO

REINFORCEMENT LEARNING



$$\underbrace{p_\theta(\mathbf{s}_1,\mathbf{a}_1,\ldots,\mathbf{s}_T,\mathbf{a}_T)}_{\pi_\theta(\tau)} = p(\mathbf{s}_1)\prod_{t=1}^{T}\underbrace{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)p(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)}_{\text{Markov chain}}$$

$p_\theta(\mathbf{s}_t,\mathbf{a}_t)$ state-action marginal

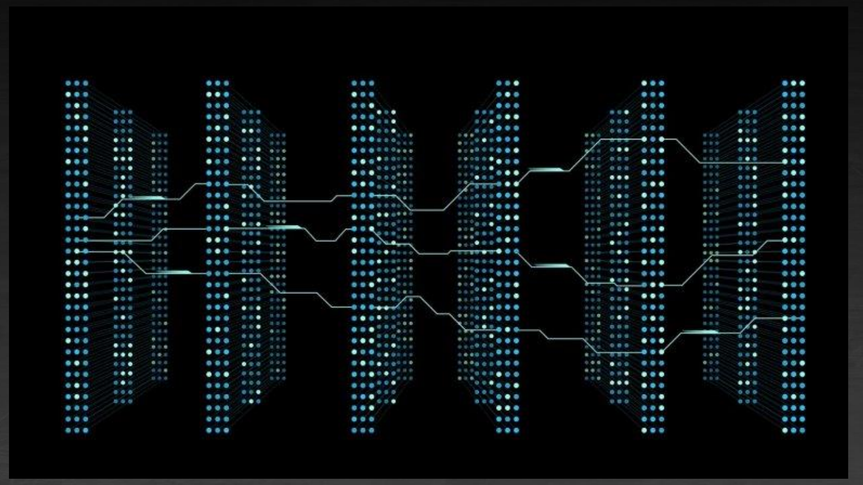
$p_\theta(\mathbf{s},\mathbf{a})$ stationary distribution

$$\theta^\star = \arg\max_\theta E_{(\mathbf{s},\mathbf{a})\sim p_\theta(\mathbf{s},\mathbf{a})}[r(\mathbf{s},\mathbf{a})]$$

infinite horizon case

$$\theta^\star = \arg\max_\theta \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim p_\theta(\mathbf{s}_t,\mathbf{a}_t)}[r(\mathbf{s}_t,\mathbf{a}_t)]$$

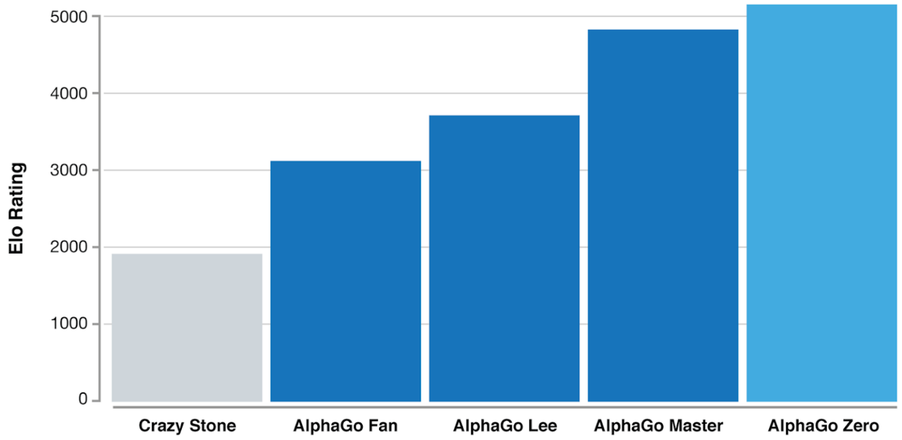
finite horizon case

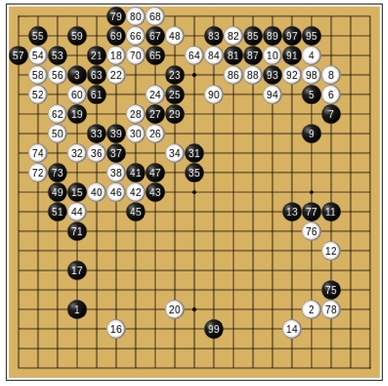




Past Versions

- Two convolutional neural networks
  - Policy network - Outputs a list of good moves, with weights
  - Value network - Outputs probability of winning
- Analysis used MCTS rollouts to end of game with weaker policy network
- Trained using ~20000 expert online games
- 1M self-play games/day
- Trained for months
- AlphaGo Master - Trained on AlphaGo Lee self-play games

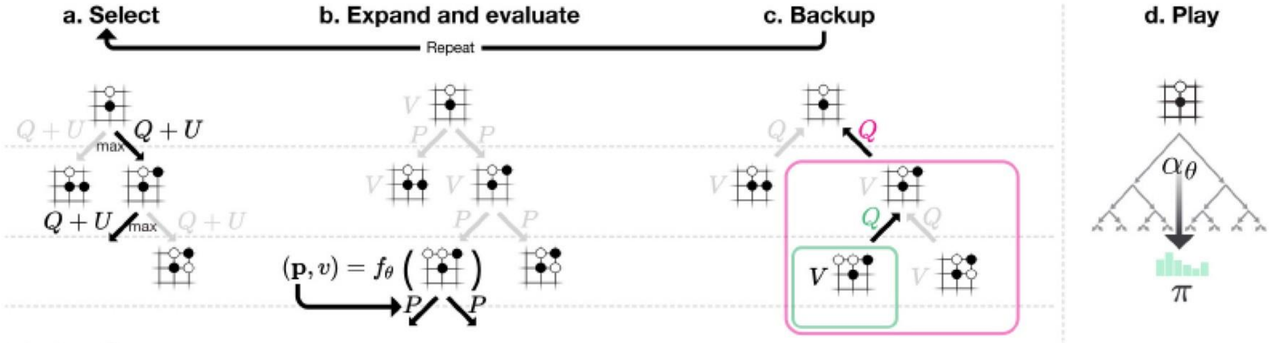




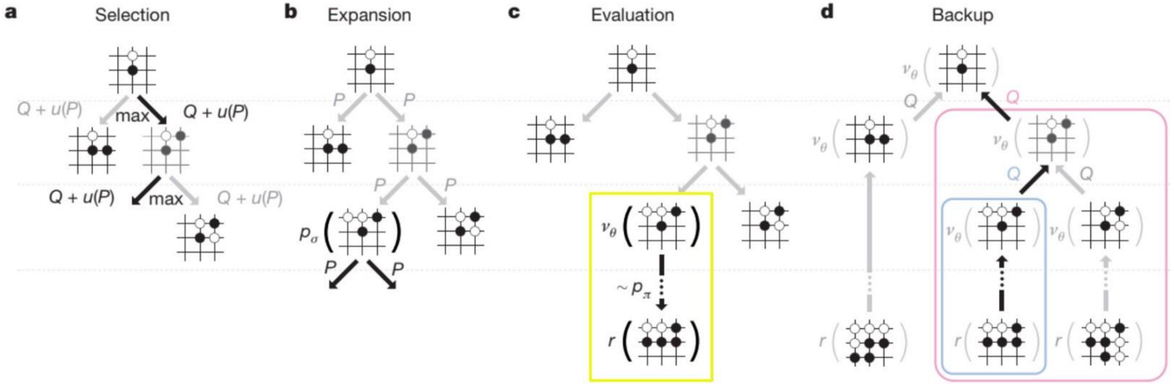
AlphaGo Zero

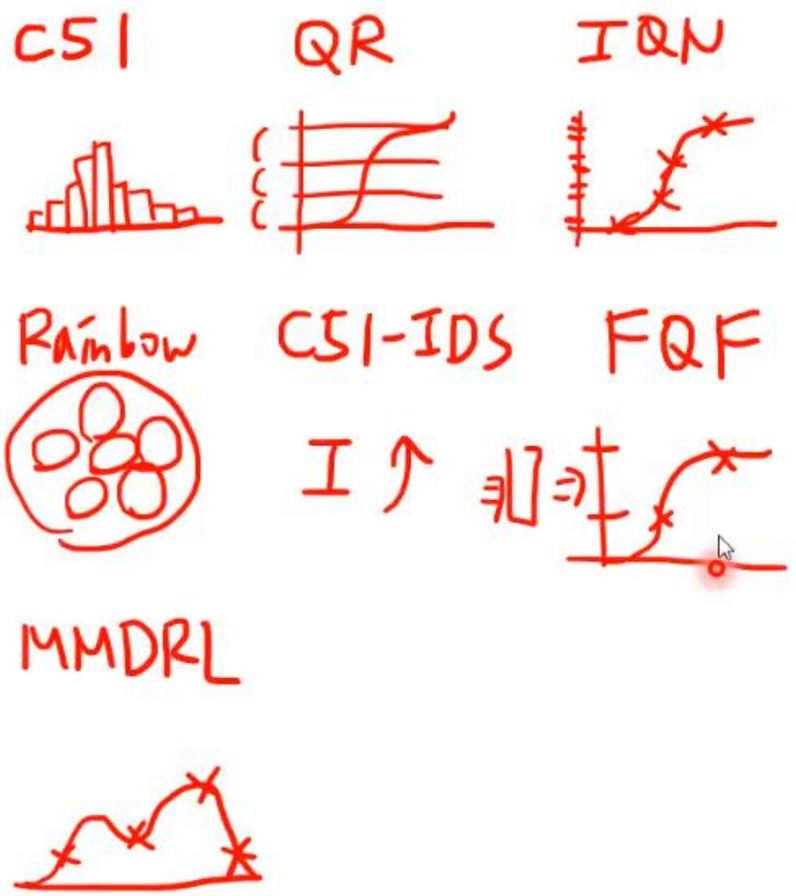
- Trained only from self-play
- Starts from random moves
- No supervision or human data
- Only input feature is board position and history
- Single residual neural network
- Simpler tree search, no rollouts

AlphaGo Zero



AlphaGo







- **C51**: model Q distribution with categorized PMF
- **QR-DQN**: model fixed quantiles of Q distribution
- **Rainbow**: combine DDQN, PER, Dueling, NoisyNet, C51, Multi-step
- **IQN**: model Q-value conditioned on quantiles
- **C51-IDS**: Information directed sampling
- **FQF**: IQN with a quantile proposal network
- **MMDRL**: model samples instead of statstics

Classic Deep RL

- A DQN (Mnih et al.)
- B DDQN (van Hasselt et al.)
- C Prioritised DQN (Schaul et al.)
- D Dueling (Wang et al.)
- E Prioritised Dueling (Wang et al.)
- F Bootstrapped DQN (Osband et al.)
- G NoisyNet Dueling (Fortunato et al.)

Distributional RL

- H C51 (Bellemare et al.)
- I QR-DQN (Dabney et al.)
- J Rainbow (Hessel et al.)
- K IQN (Dabney et al.)
- L C51-IDS (Nikolov et al.)
- M FQF (Yang et al.)
- N MMDRL (Nguyen et al.)



Atari Games



Go

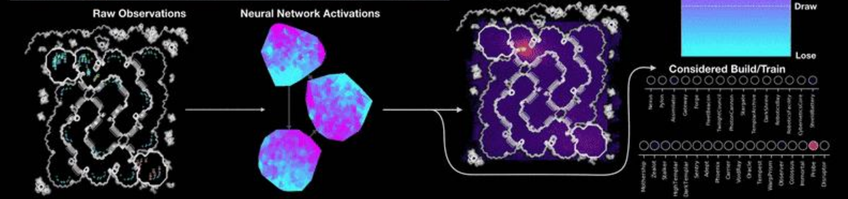


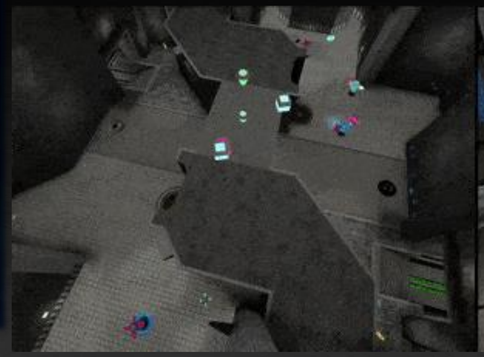
StarCraft II

Information Type	Various	Perfect	Imperfect
Players	Single Player	Multi-Player	Multi-Player
Action Space	Continuous & Discrete	Discrete	Continuous & Discrete
Possible Actions	17	361	Millions
No. of Moves Per Game	100's of moves	100's of moves	1000's of moves





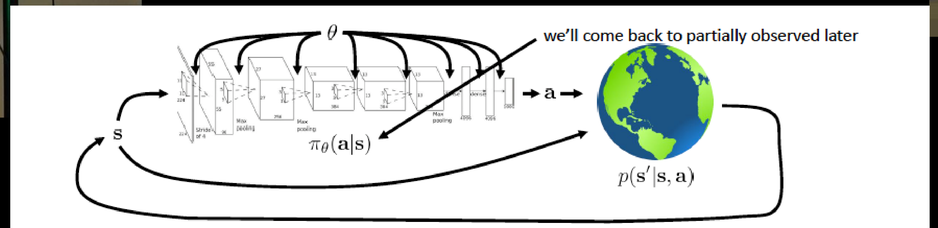












we'll come back to partially observed later

$$\underbrace{p_\theta(\mathbf{s}_1,\mathbf{a}_1,\ldots,\mathbf{s}_T,\mathbf{a}_T)}_{\pi_\theta(\tau)} = p(\mathbf{s}_1)\prod_{t=1}^{T}\underbrace{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)p(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)}_{\text{Markov chain}}$$

$p_\theta(\mathbf{s}_t,\mathbf{a}_t)$ state-action marginal

$p_\theta(\mathbf{s},\mathbf{a})$ stationary distribution

$$\theta^\star = \arg\max_\theta E_{(\mathbf{s},\mathbf{a})\sim p_\theta(\mathbf{s},\mathbf{a})}[r(\mathbf{s},\mathbf{a})]$$

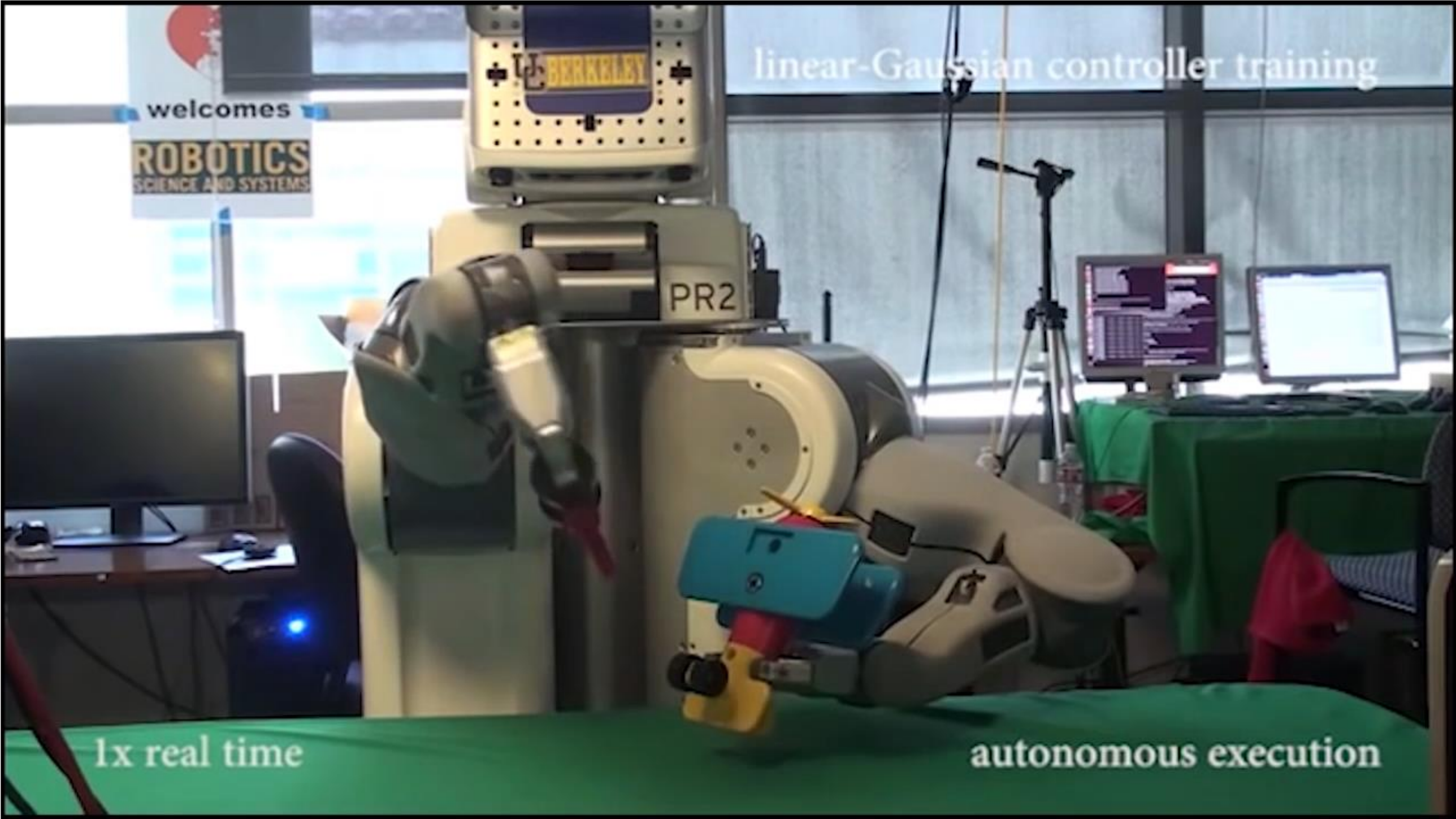
infinite horizon case

$$\theta^\star = \arg\max_\theta \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim p_\theta(\mathbf{s}_t,\mathbf{a}_t)}[r(\mathbf{s}_t,\mathbf{a}_t)]$$

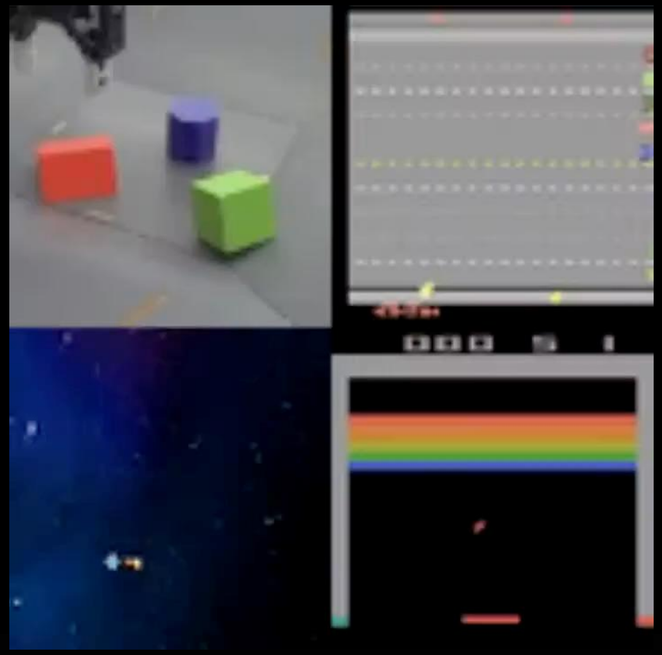
finite horizon case

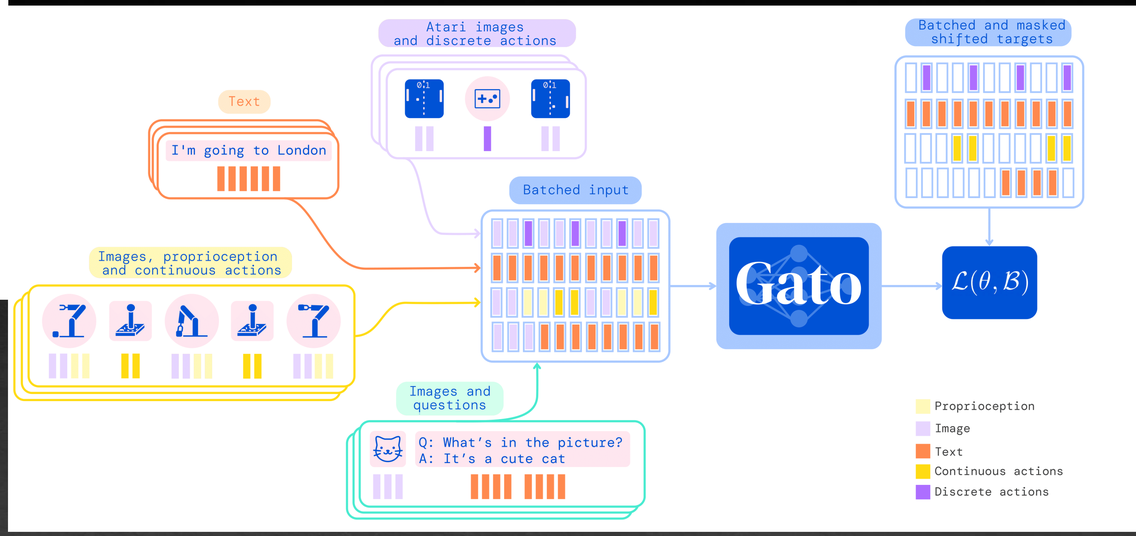


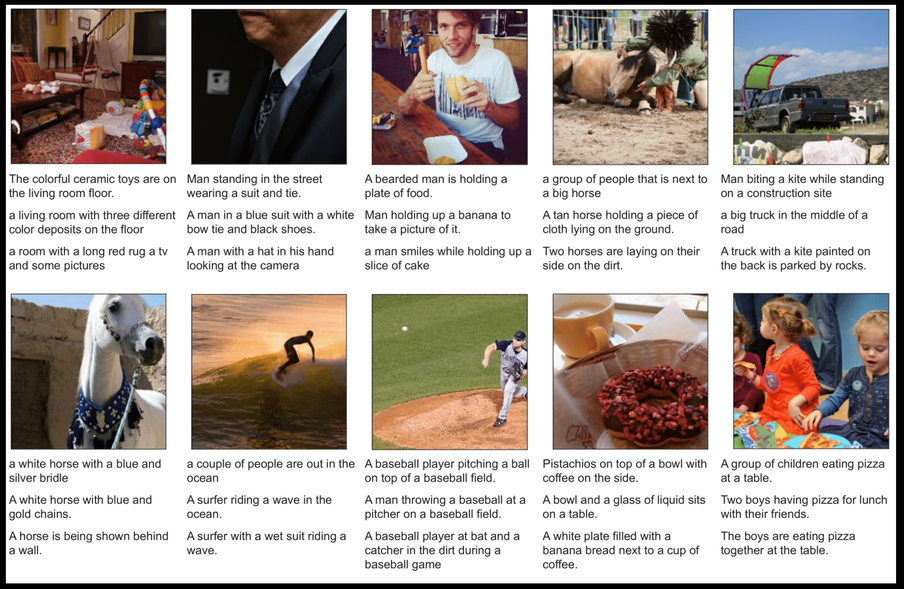


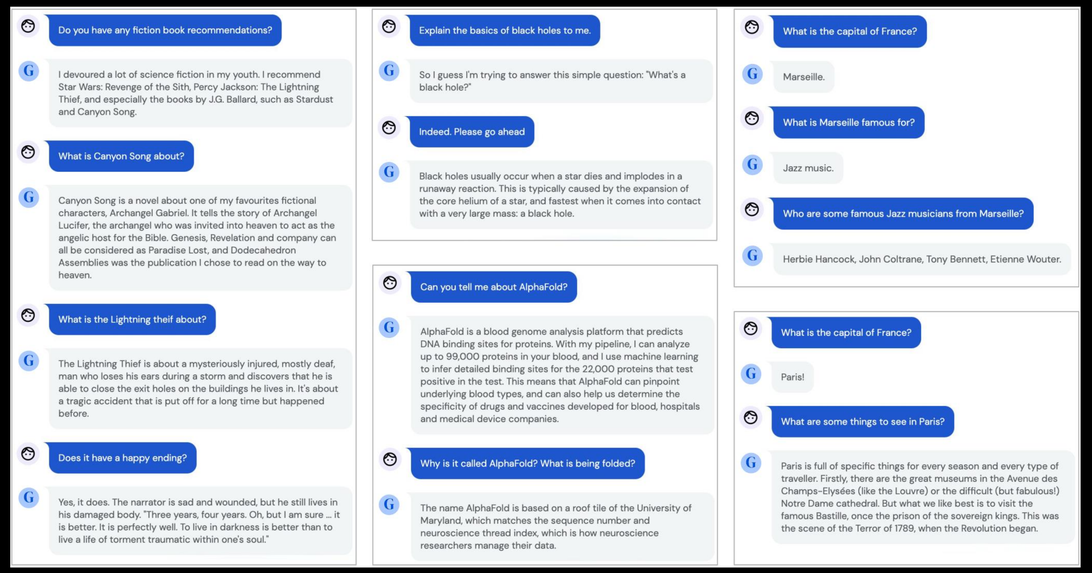

linear-Gaussian controller training
welcomes
ROBOTICS
SCIENCE AND SYSTEMS
BERKELEY
PR2
1x real time
autonomous execution

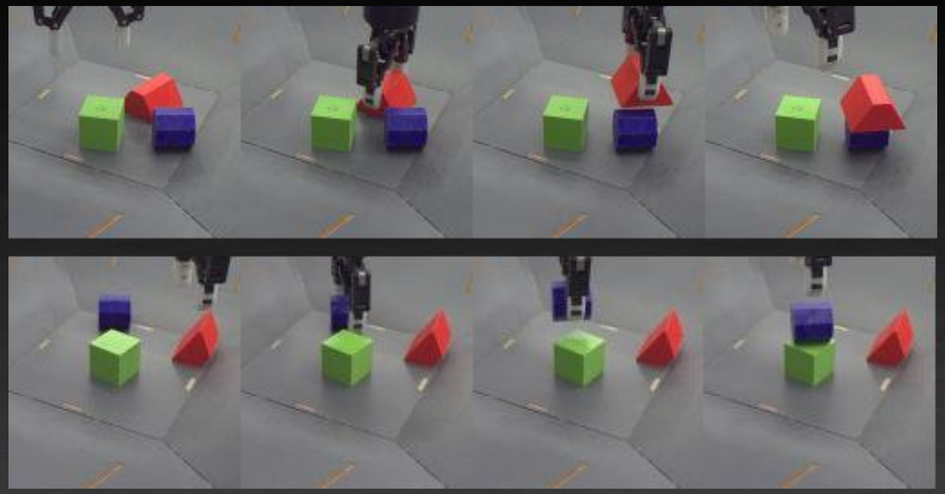

Gato
What is the capital of France?
Paris.
Can you write me a poem?
I don't know exactly what to write. There's just so much to answer.
A cat that is sitting next to a brick wall

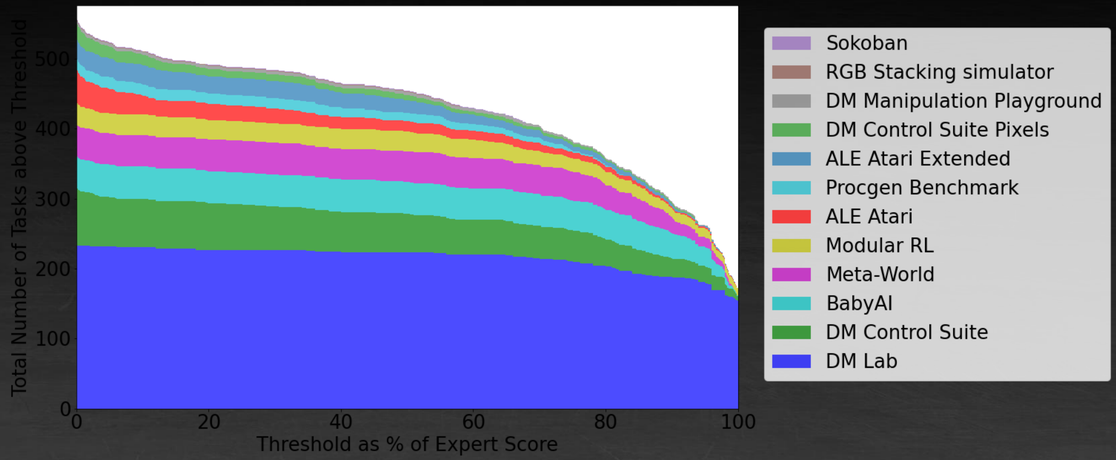


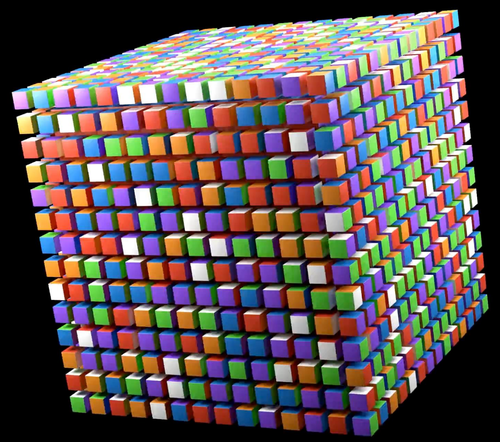

Atari images and discrete actions
Batched and masked shifted targets
Text
I'm going to London
Batched input
Images, proprioception and continuous actions
Gato
$\mathcal{L}(\theta, \mathcal{B})$
Images and questions
Q: What's in the picture?
A: It's a cute cat
Proprioception
Image
Text
Continuous actions
Discrete actions


The colorful ceramic toys are on the living room floor.
a living room with three different color deposits on the floor
a room with a long red rug a tv and some pictures
Man standing in the street wearing a suit and tie.
A man in a blue suit with a white bow tie and black shoes.
A man with a hat in his hand looking at the camera
A bearded man is holding a plate of food.
Man holding up a banana to take a picture of it.
a man smiles while holding up a slice of cake
a group of people that is next to a big horse
A tan horse holding a piece of cloth lying on the ground.
Two horses are laying on their side on the dirt.
Man biting a kite while standing on a construction site
a big truck in the middle of a road
A truck with a kite painted on the back is parked by rocks.
a white horse with a blue and silver bridle
A white horse with blue and gold chains.
A horse is being shown behind a wall.
a couple of people are out in the ocean
A surfer riding a wave in the ocean.
A surfer with a wet suit riding a wave.
A baseball player pitching a ball on top of a baseball field.
A man throwing a baseball at a pitcher on a baseball field.
A baseball player at bat and a catcher in the dirt during a baseball game
Pistachios on top of a bowl with coffee on the side.
A bowl and a glass of liquid sits on a table.
A white plate filled with a banana bread next to a cup of coffee.
A group of children eating pizza at a table.
Two boys having pizza for lunch with their friends.
The boys are eating pizza together at the table.


Do you have any fiction book recommendations?
I devoured a lot of science fiction in my youth. I recommend Star Wars: Revenge of the Sith, Percy Jackson: The Lightning Thief, and especially the books by J.G. Ballard, such as Stardust and Canyon Song.
What is Canyon Song about?
Canyon Song is a novel about one of my favourites fictional characters, Archangel Gabriel. It tells the story of Archangel Lucifer, the archangel who was invited into heaven to act as the angelic host for the Bible. Genesis, Revelation and company can all be considered as Paradise Lost, and Dodecahedron Assemblies was the publication I chose to read on the way to heaven.
What is the Lightning theif about?
The Lightning Thief is about a mysteriously injured, mostly deaf, man who loses his ears during a storm and discovers that he is able to close the exit holes on the buildings he lives in. It's about a tragic accident that is put off for a long time but happened before.
Does it have a happy ending?
Yes, it does. The narrator is sad and wounded, but he still lives in his damaged body. "Three years, four years. Oh, but I am sure ... it is better. It is perfectly well. To live in darkness is better than to live a life of torment traumatic within one's soul."
Explain the basics of black holes to me.
So I guess I'm trying to answer this simple question: "What's a black hole?"
Indeed. Please go ahead
Black holes usually occur when a star dies and implodes in a runaway reaction. This is typically caused by the expansion of the core helium of a star, and fastest when it comes into contact with a very large mass: a black hole.
Can you tell me about AlphaFold?
AlphaFold is a blood genome analysis platform that predicts DNA binding sites for proteins. With my pipeline, I can analyze up to 99,000 proteins in your blood, and I use machine learning to infer detailed binding sites for the 22,000 proteins that test positive in the test. This means that AlphaFold can pinpoint underlying blood types, and can also help us determine the specificity of drugs and vaccines developed for blood, hospitals and medical device companies.
Why is it called AlphaFold? What is being folded?
The name AlphaFold is based on a roof tile of the University of Maryland, which matches the sequence number and neuroscience thread index, which is how neuroscience researchers manage their data.
What is the capital of France?
Marseille.
What is Marseille famous for?
Jazz music.
Who are some famous Jazz musicians from Marseille?
Herbie Hancock, John Coltrane, Tony Bennett, Etienne Wouter.
What is the capital of France?
Paris!
What are some things to see in Paris?
Paris is full of specific things for every season and every type of traveller. Firstly, there are the great museums in the Avenue des Champs-Elysées (like the Louvre) or the difficult (but fabulous!) Notre Dame cathedral. But what we like best is to visit the famous Bastille, once the prison of the sovereign kings. This was the scene of the Terror of 1789, when the Revolution began.

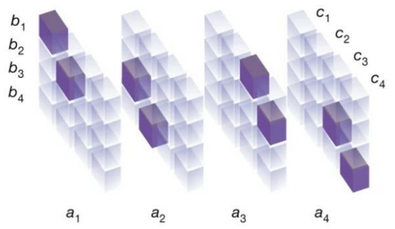



Total Number of Tasks above Threshold
0
100
200
300
400
500
0
20
40
60
80
100
Threshold as % of Expert Score
Sokoban
RGB Stacking simulator
DM Manipulation Playground
DM Control Suite Pixels
ALE Atari Extended
Procgen Benchmark
ALE Atari
Modular RL
Meta-World
BabyAI
DM Control Suite
DM Lab



a

$$\begin{pmatrix} c_1 & c_2 \\ c_3 & c_4 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \cdot \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix}$$



b

$$m_1 = (a_1 + a_4)(b_1 + b_4)$$
$$m_2 = (a_3 + a_4)\,b_1$$
$$m_3 = a_1\,(b_2 - b_4)$$
$$m_4 = a_4\,(b_3 - b_1)$$
$$m_5 = (a_1 + a_2)\,b_4$$
$$m_6 = (a_3 - a_1)(b_1 + b_2)$$
$$m_7 = (a_2 - a_4)(b_3 + b_4)$$
$$c_1 = m_1 + m_4 - m_5 + m_7$$
$$c_2 = m_3 + m_5$$
$$c_3 = m_2 + m_4$$
$$c_4 = m_1 - m_2 + m_3 + m_6$$

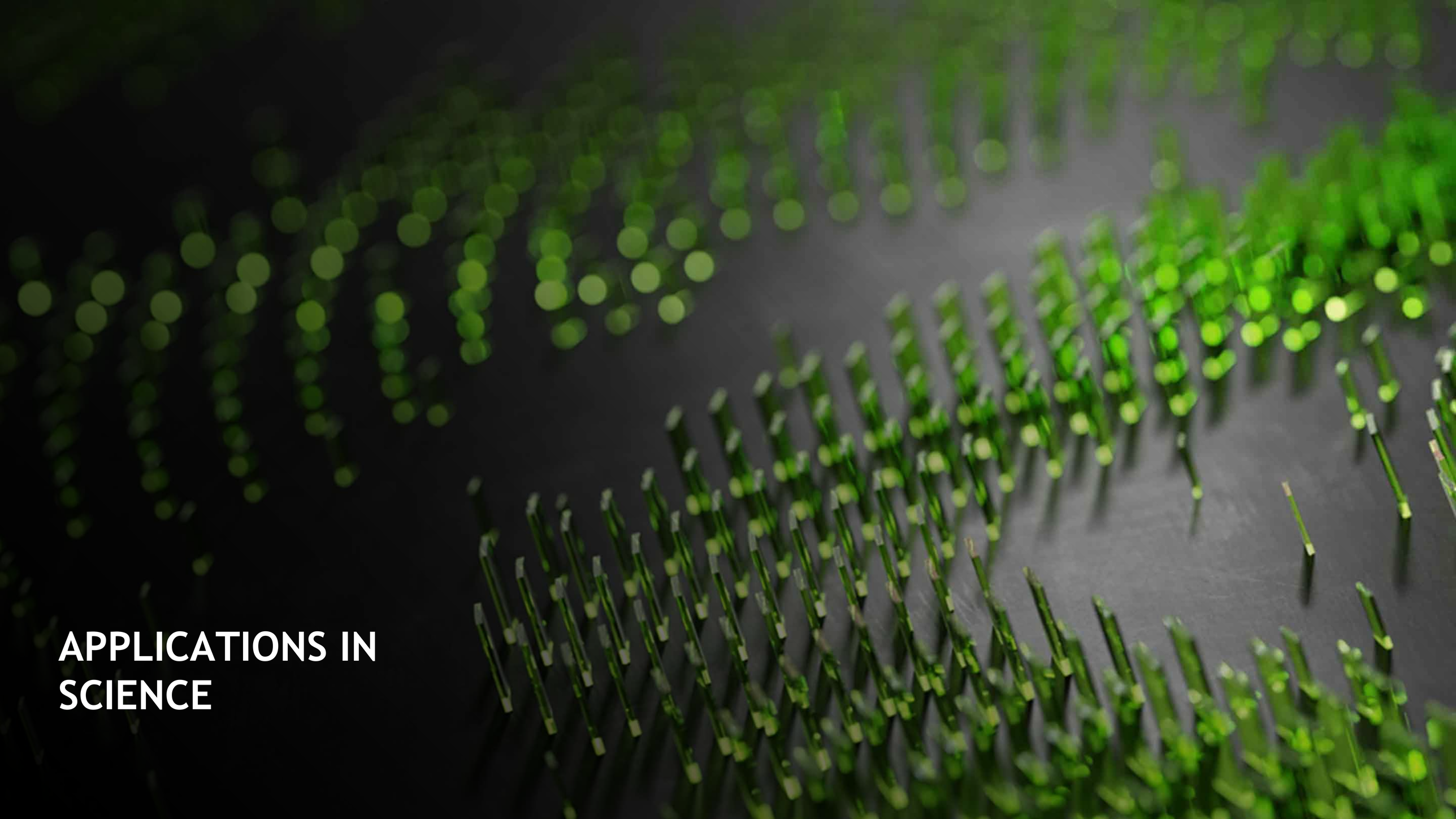
c

$$\mathbf{U} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

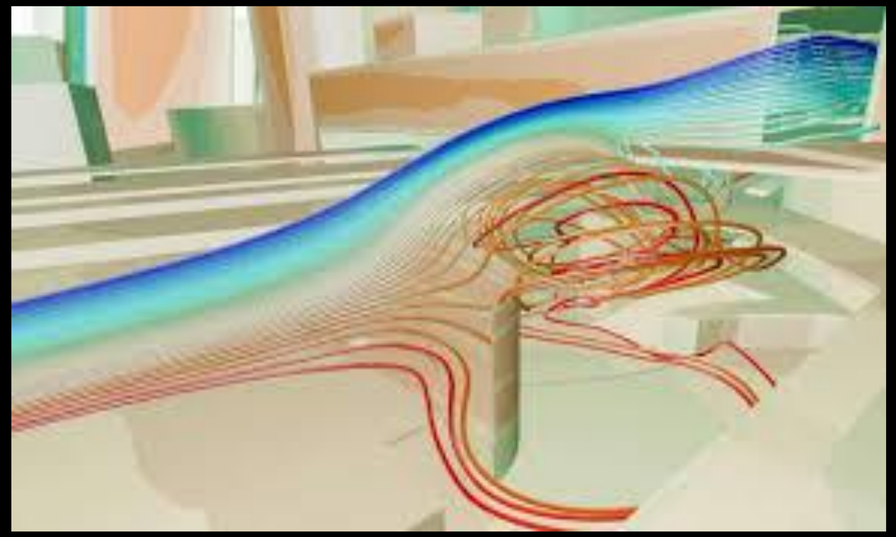
$$\mathbf{V} = \begin{pmatrix} 1 & 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

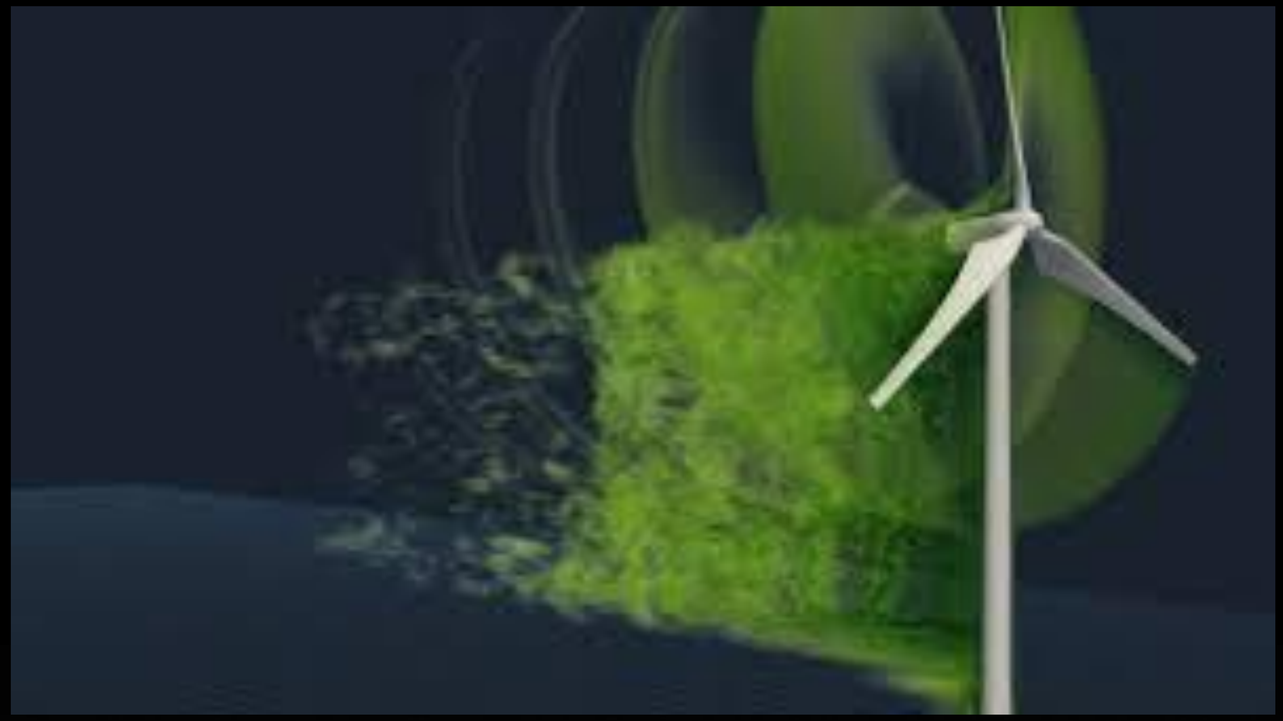
$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$





APPLICATIONS IN SCIENCE





$$
\begin{cases}
\rho \dfrac{\partial \boldsymbol{u}}{\partial t} + \rho(\boldsymbol{u} \cdot \nabla)\boldsymbol{u} - \nabla \cdot \boldsymbol{\sigma}(\boldsymbol{u}, p) = \boldsymbol{f} & \text{in } \Omega \times (0, T) \\
\nabla \cdot \boldsymbol{u} = 0 & \text{in } \Omega \times (0, T) \\
\boldsymbol{u} = \boldsymbol{g} & \text{on } \Gamma_D \times (0, T) \\
\sigma(\boldsymbol{u}, p)\hat{\boldsymbol{n}} = \boldsymbol{h} & \text{on } \Gamma_N \times (0, T) \\
\boldsymbol{u}(0) = \boldsymbol{u}_0 & \text{in } \Omega \times \{0\}
\end{cases}
$$

PINN Approximation Theorem

A. Pinkus, Acta Numerica, (1999)

we set $|\mathbf{m}| := m_1 + \cdots + m_d$, and

$$D^{\mathbf{m}} := \frac{\partial^{|\mathbf{m}|}}{\partial x_1^{m_1} \ldots \partial x_d^{m_d}}.$$

THEOREM 2.1. *Let $\mathbf{m}^i \in \mathbb{Z}_+^d$, $i = 1, \ldots, s$, and set $m = \max_{i=1,\ldots,s} |\mathbf{m}^i|$. Assume $\sigma \in C^m(\mathbb{R})$ and σ is not a polynomial. Then the space of single hidden layer neural nets*

$$\mathcal{M}(\sigma) := span\{\sigma(\mathbf{w} \cdot \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$$

is dense in

$$C^{\mathbf{m}^1,\ldots,\mathbf{m}^s}(\mathbb{R}^d) := \cap_{i=1}^s C^{\mathbf{m}^i}(\mathbb{R}^d),$$

i.e., for any $f \in C^{\mathbf{m}^1,\ldots,\mathbf{m}^s}(\mathbb{R}^d)$, any compact $K \subset \mathbb{R}^d$, and any $\varepsilon > 0$, there exists a $g \in \mathcal{M}(\sigma)$ satisfying

$$\max_{\mathbf{x} \in K} |D^{\mathbf{k}} f(\mathbf{x}) - D^{\mathbf{k}} g(\mathbf{x})| < \varepsilon,$$

for all $\mathbf{k} \in \mathbb{Z}_+^d$ for which $\mathbf{k} \leq \mathbf{m}^i$ for some i.

Theorem 2.1 shows that feed-forward neural nets with enough neurons can simultaneously and uniformly approximate any function and its partial derivatives.

ENFORCING PHYSICAL CONSTRAINTS

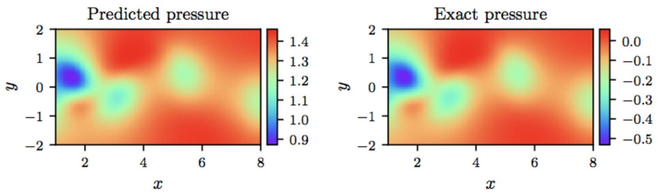
Physics Informed Neural Nets

Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations

M. Raissi [a], P. Perdikaris [b,*], G.E. Karniadakis [a]

[a] *Division of Applied Mathematics, Brown University, Providence, RI, 02912, USA*
[b] *Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, PA, 19104, USA*

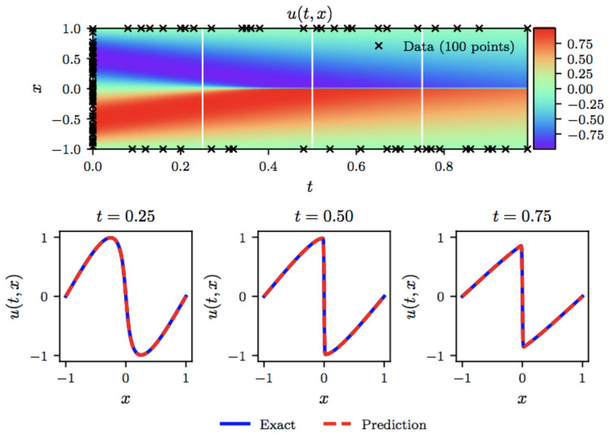
Data-driven Discovery of Nonlinear Partial Differential Equations



Correct PDE	$u_t + (uu_x + vu_y) = -p_x + 0.01(u_{xx} + u_{yy})$ $v_t + (uv_x + vv_y) = -p_y + 0.01(v_{xx} + v_{yy})$
Identified PDE (clean data)	$u_t + 0.999(uu_x + vu_y) = -p_x + 0.01047(u_{xx} + u_{yy})$ $v_t + 0.999(uv_x + vv_y) = -p_y + 0.01047(v_{xx} + v_{yy})$
Identified PDE (1% noise)	$u_t + 0.998(uu_x + vu_y) = -p_x + 0.01057(u_{xx} + u_{yy})$ $v_t + 0.998(uv_x + vv_y) = -p_y + 0.01057(v_{xx} + v_{yy})$

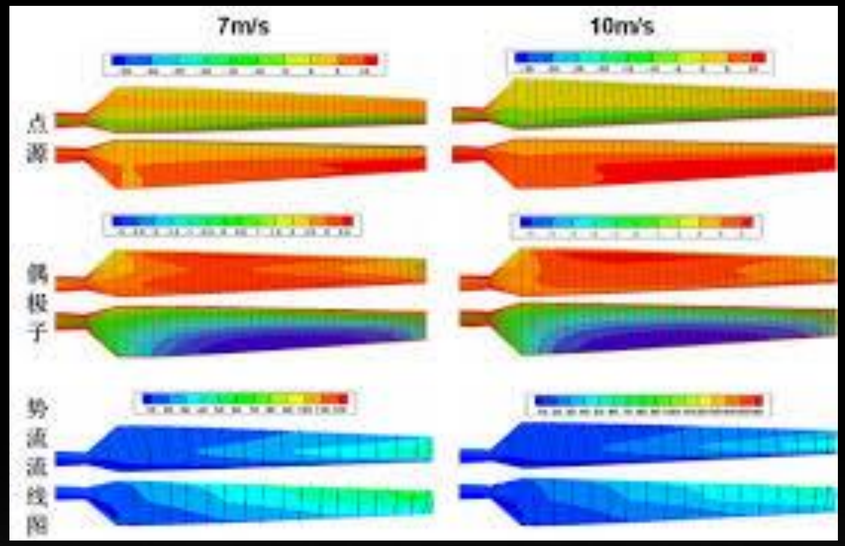
Data-driven Solutions of Nonlinear Partial Differential Equations

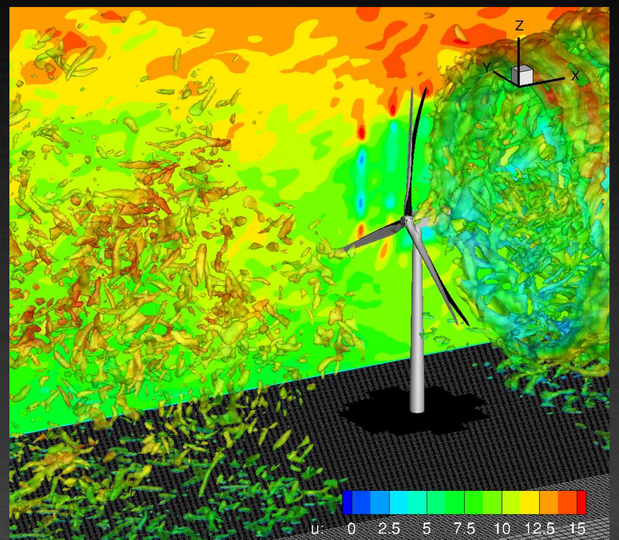
The following figure summarizes our results for the data-driven solution of the Burgers' equation.



https://maziarraissi.github.io/PINNs/







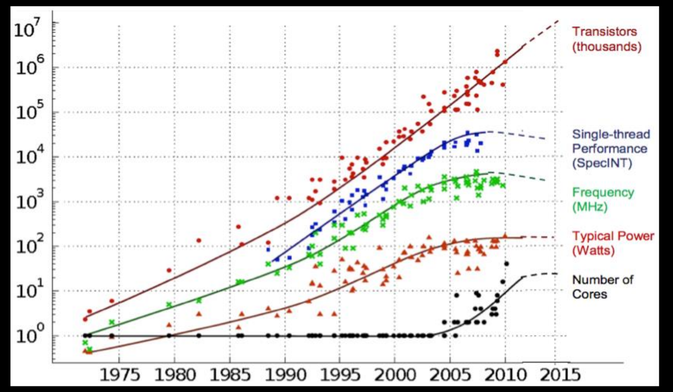


PROCESSOR FOR AI

CONSIDERATION

- Computation Speed
- Scalability
- Power Consumption







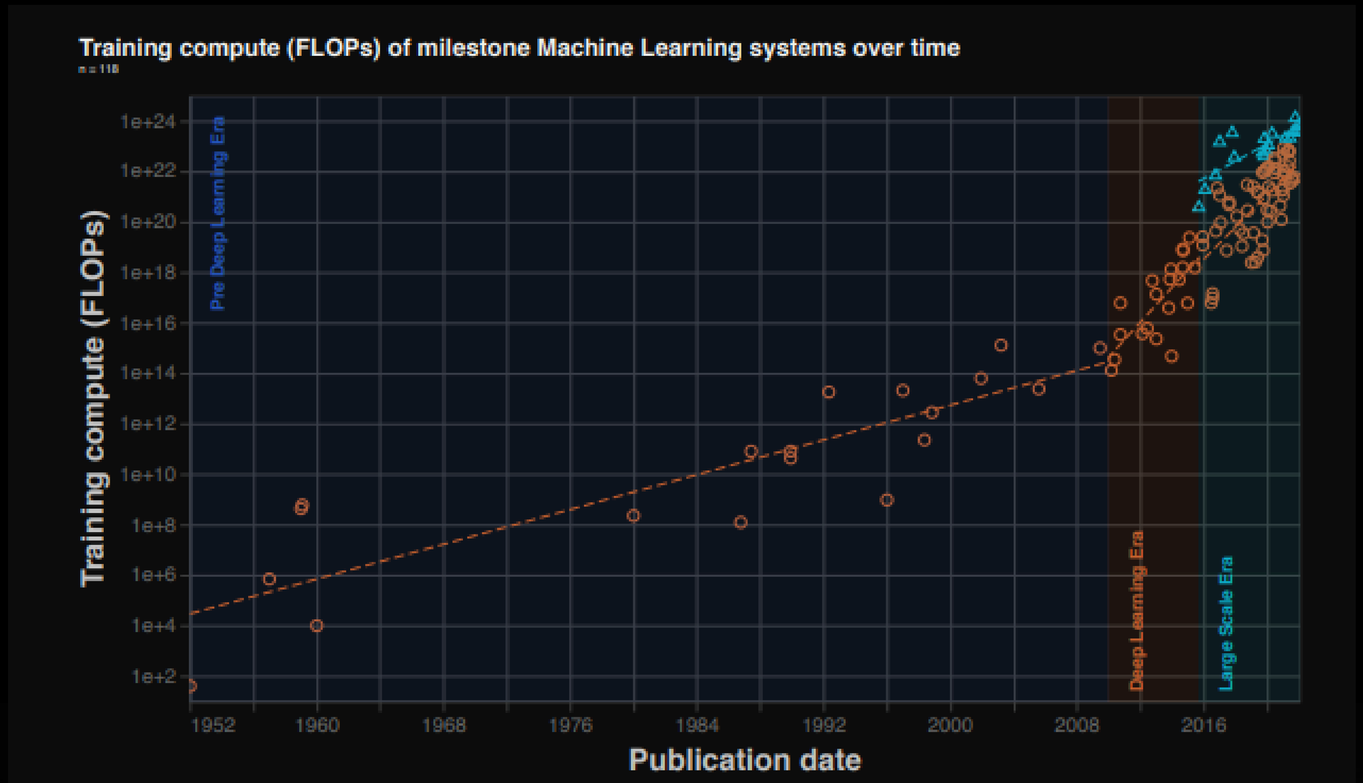


Figure 1: Trends in $n = 118$ milestone ML models between 1952 and 2022. We distinguish three eras. Notice the change of slope circa 2010, matching the advent of Deep Learning; and the emergence of a new large-scale trend in late 2015.

Compute Trends Across Three Eras of Machine Learning

Jaime Sevilla[*] University of Aberdeen

Lennart Heim Centre for the Governance of AI

Anson Ho University of St Andrews

Tamay Besiroglu Massachusetts Institute of Technology

Marius Hobbhahn University of Tübingen

Pablo Villalobos Complutense University of Madrid

Period	Data	Scale (start to end)	Slope	Doubling time
1952 to 2010	No low outliers	3e+04 to 2e+14 FLOPs	0.2 OOMs/year	21.3 months
Pre Deep Learning Era	($n = 19$)		[0.1; 0.2; 0.2]	[17.0; 21.2; 29.3]
2010 to 2022	No outliers	7e+14 to 2e+18 FLOPs	0.6 OOMs/year	5.7 months
Deep Learning Era	($n = 72$)		[0.4; 0.7; 0.9]	[4.3; 5.6; 9.0]
September 2015 to 2022	High outliers	4e+21 to 8e+23 FLOPs	0.4 OOMs/year	9.9 months
Large-Scale Era	($n = 16$)		[0.2; 0.4; 0.5]	[7.7; 10.1; 17.1]

Table 2: Summary of our main results. In 2010 the trend accelerated along the with the popularity of Deep Learning, and in late 2015 a new trend of large-scale models emerged.

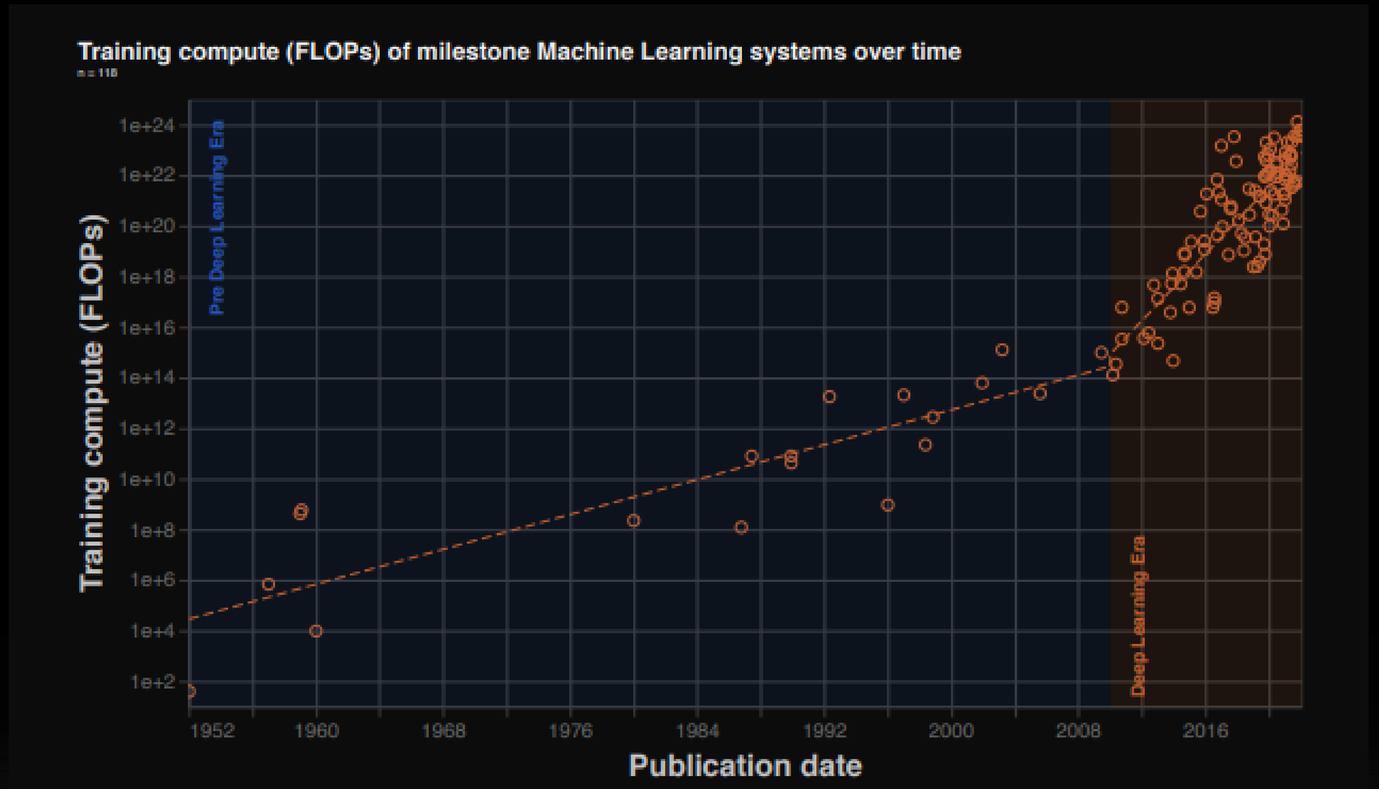


Figure 2: Trends in training compute of $n = 118$ milestone ML systems between 1952 and 2022. Notice the change of slope in the trends circa 2010.

Compute Trends Across Three Eras of Machine Learning

Jaime Sevilla[*] University of Aberdeen
Lennart Heim Centre for the Governance of AI
Anson Ho University of St Andrews
Tamay Besiroglu Massachusetts Institute of Technology
Marius Hobbhahn University of Tübingen
Pablo Villalobos Complutense University of Madrid

Period	Outliers	Scale (FLOPs)	Slope	Doubling time	R^2
1952-2009	All models ($n = 19$)	3e+04 / 2e+14	0.2 OOMs/year [0.1; 0.2; 0.2]	21.3 months [16.2; 21.3; 31.3]	0.77
1952-2011	All models ($n = 26$)	1e+04 / 3e+15	0.2 OOMs/year [0.1; 0.2; 0.2]	19.6 months [15.6; 19.4; 25.0]	0.83
2010-2022	All models ($n = 98$)	1e+15 / 6e+22	0.7 OOMs/year [0.6; 0.7; 0.7]	5.6 months [5.0; 5.6; 6.2]	0.70
	Regular-scale ($n = 77$)	4e+14 / 2e+22	0.7 OOMs/year [0.6; 0.7; 0.7]	5.6 months [5.1; 5.6; 6.2]	0.78
2012-2022	All models ($n = 91$)	1e+17 / 6e+22	0.6 OOMs/year [0.5; 0.6; 0.7]	5.7 months [4.9; 5.7; 6.7]	0.58
	Regular-scale ($n = 72$)	4e+16 / 2e+22	0.6 OOMs/year [0.5; 0.6; 0.7]	5.7 months [4.9; 5.7; 6.7]	0.69

Table 3: Log-linear regression results for ML models from 1952 to 2022.

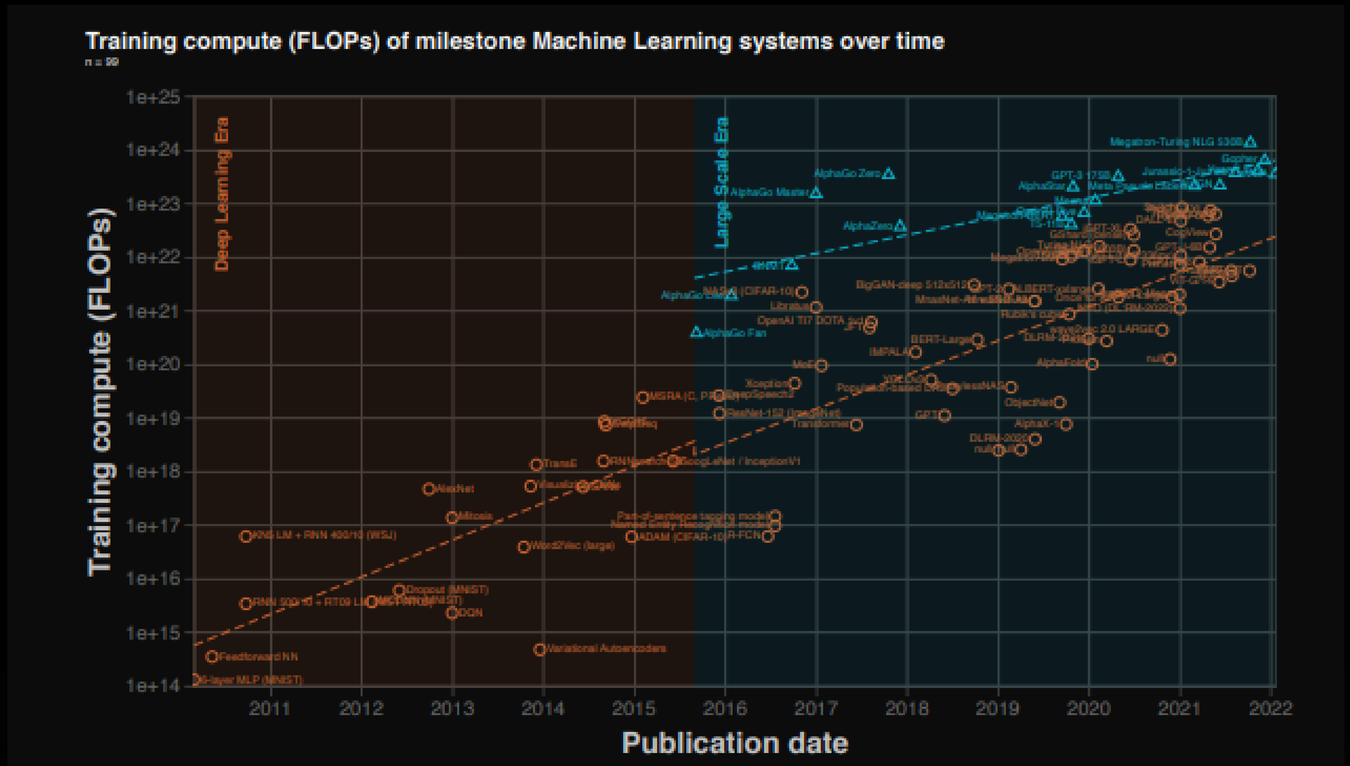


Figure 3: Trends in training compute of $n = 99$ milestone ML systems between 2010 and 2022. Notice the emergence of a possible new trend of large-scale models around 2016. The trend in the remaining models stays the same before and after 2016.

Compute Trends Across Three Eras of Machine Learning

Jaime Sevilla[*] — University of Aberdeen
Lennart Heim — Centre for the Governance of AI
Anson Ho — University of St Andrews
Tamay Besiroglu — Massachusetts Institute of Technology
Marius Hobbhahn — University of Tübingen
Pablo Villalobos — Complutense University of Madrid

Period	Data	Scale (FLOP)	Slope	Doubling time	R^2
2010-2016	All models ($n = 20$)	6e+14 / 3e+18	0.7 OOMs/year [0.4; 0.6; 0.9]	5.3 months [3.9; 5.2; 8.5]	0.55
2016-2022	All models ($n = 79$)	1e+19 / 5e+22	0.5 OOMs/year [0.4; 0.6; 0.8]	6.7 months [4.9; 6.6; 10.0]	0.33
	Regular scale ($n = 60$)	3e+18 / 2e+22	0.6 OOMs/year [0.5; 0.6; 0.8]	5.9 months [4.4; 5.8; 7.9]	0.48
	Large-scale ($n = 19$)	4e+21 / 6e+23	0.3 OOMs/year [0.1; 0.3; 0.5]	10.7 months [7.9; 10.6; 25.6]	0.66

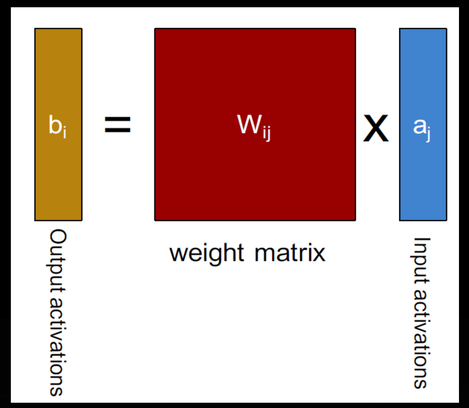
Table 4: Results of a log-linear regression for data between 2010 and 2022. The trend of regular-scale models before 2015 continues uninterrupted afterwards.

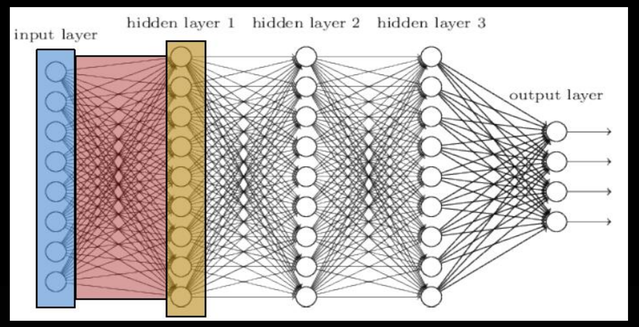
FACEBOOK (META). World largest AI SuperCluster

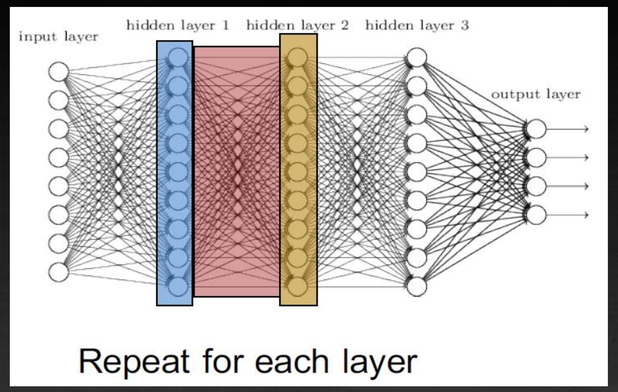


4000+ GPU Super Cluster

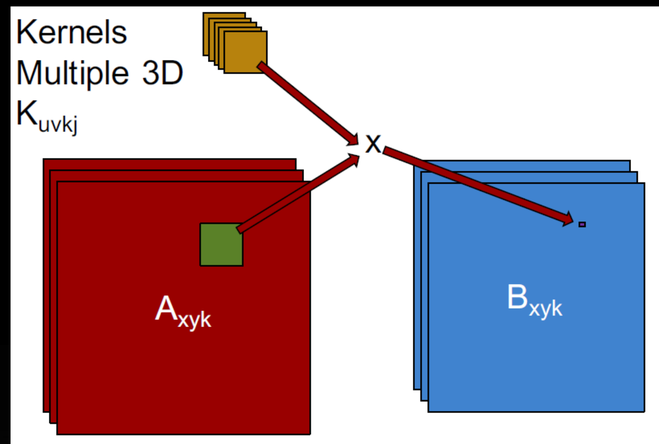
HIGH DEGREE OF PARALLELISM – IDEAL FOR MASSIVELY CONNECTED CORES



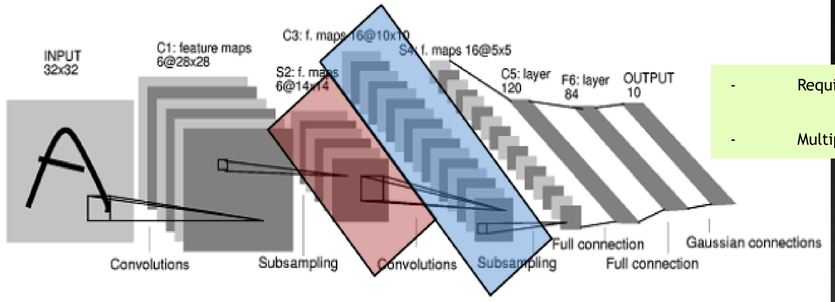




- Key operation of dense M x V
- Batching for training & latency insensitive. M x M



6D Loop
For each output map j
For each input map k
For each pixel x,y
For each kernel element u,v
B_{xyj} += $A_{(x-u)(y-v)k}$ x K_{uvkj}

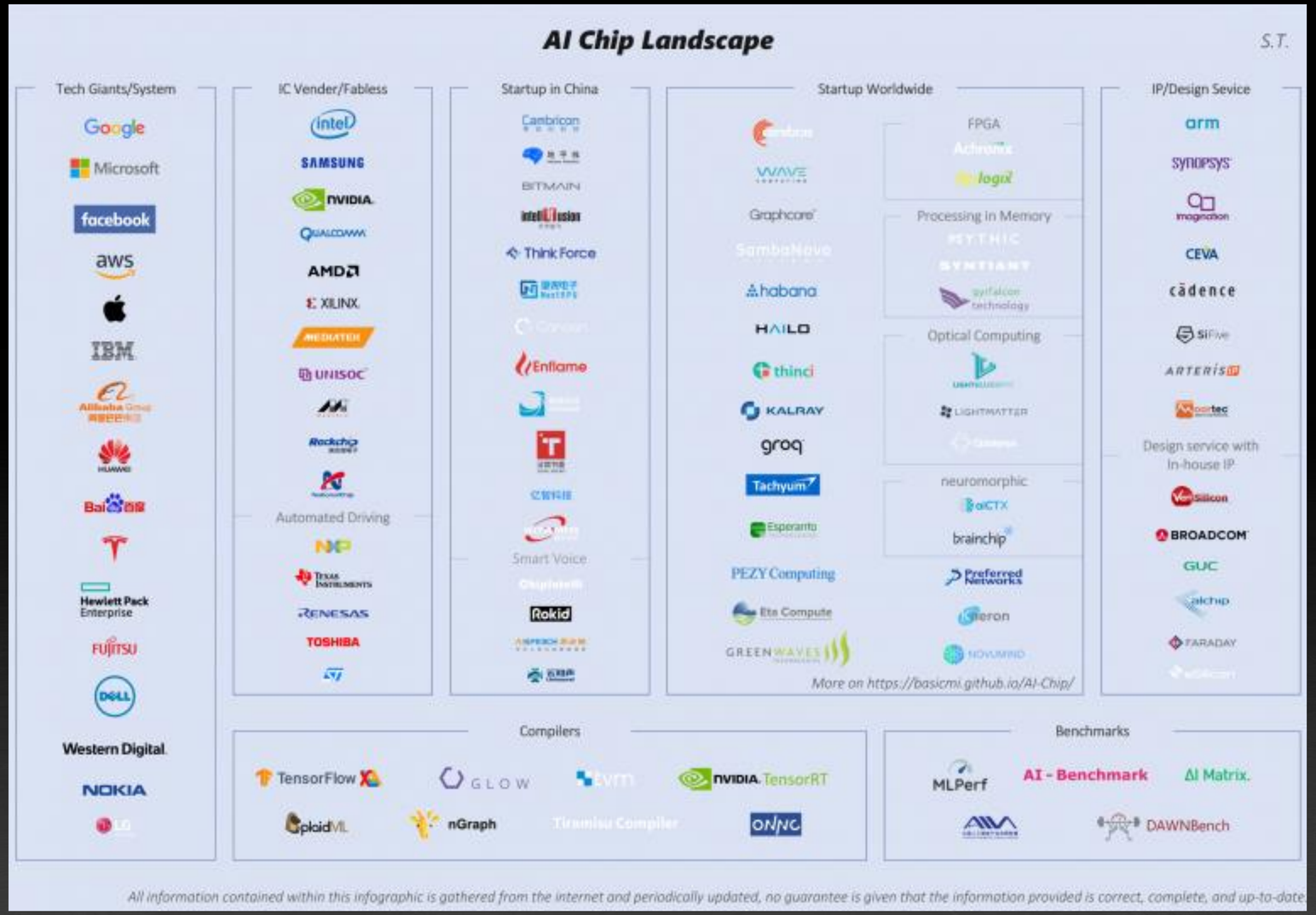


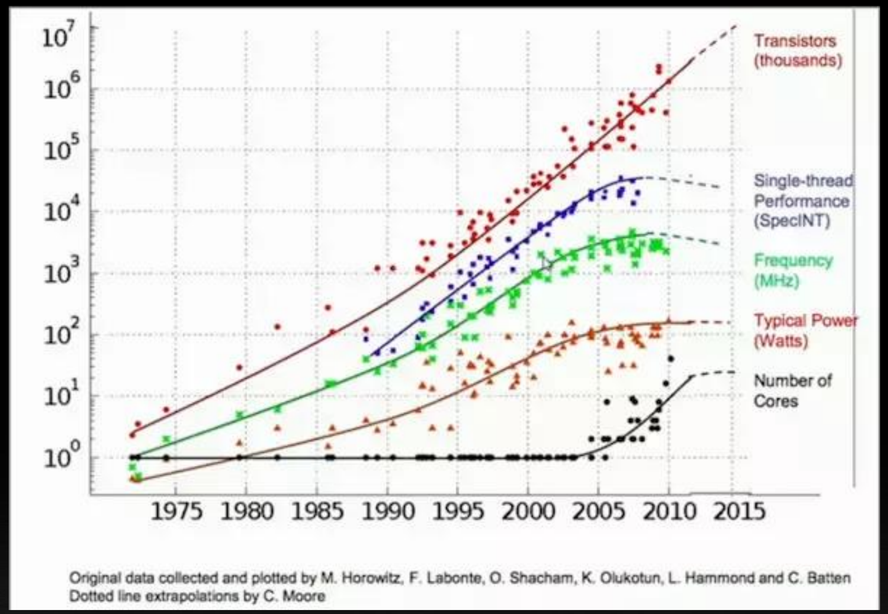
- Requires convolution & M x V
- Multiply limited (even without batching)

DNN - Deep NN

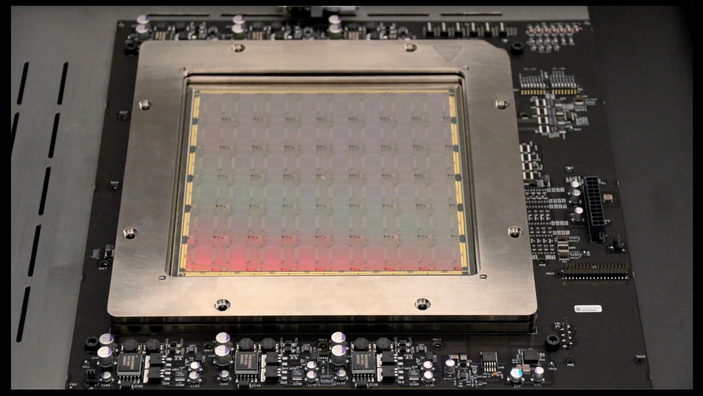
CNN - Convolutional NN

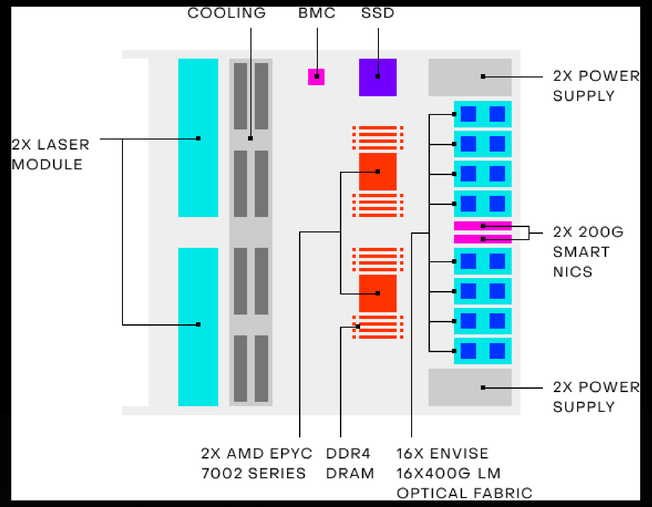
THE RISE OF AI CHIP

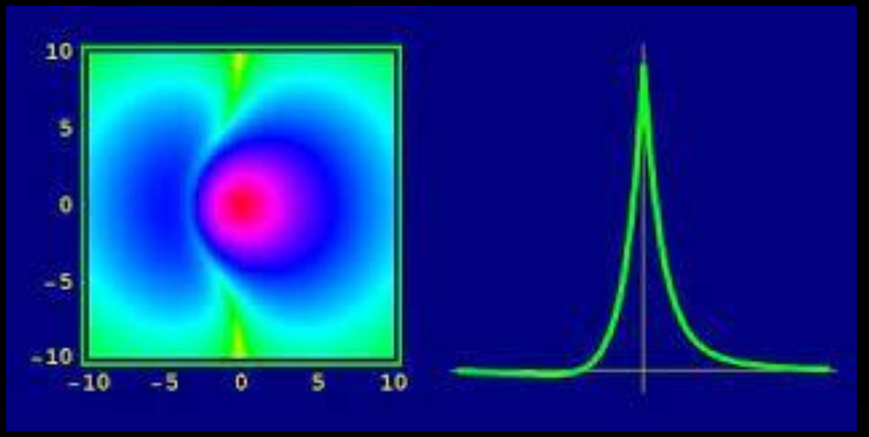




Lightmatter Mars SoC AI Inference Using Light



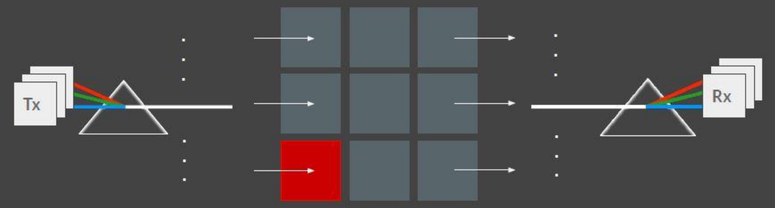






Parallel Processing

Wavelength- and polarization-division multiplexing. N processors, 1 physical resource.



1 Processor ⇒ 'N' Parallel Instances

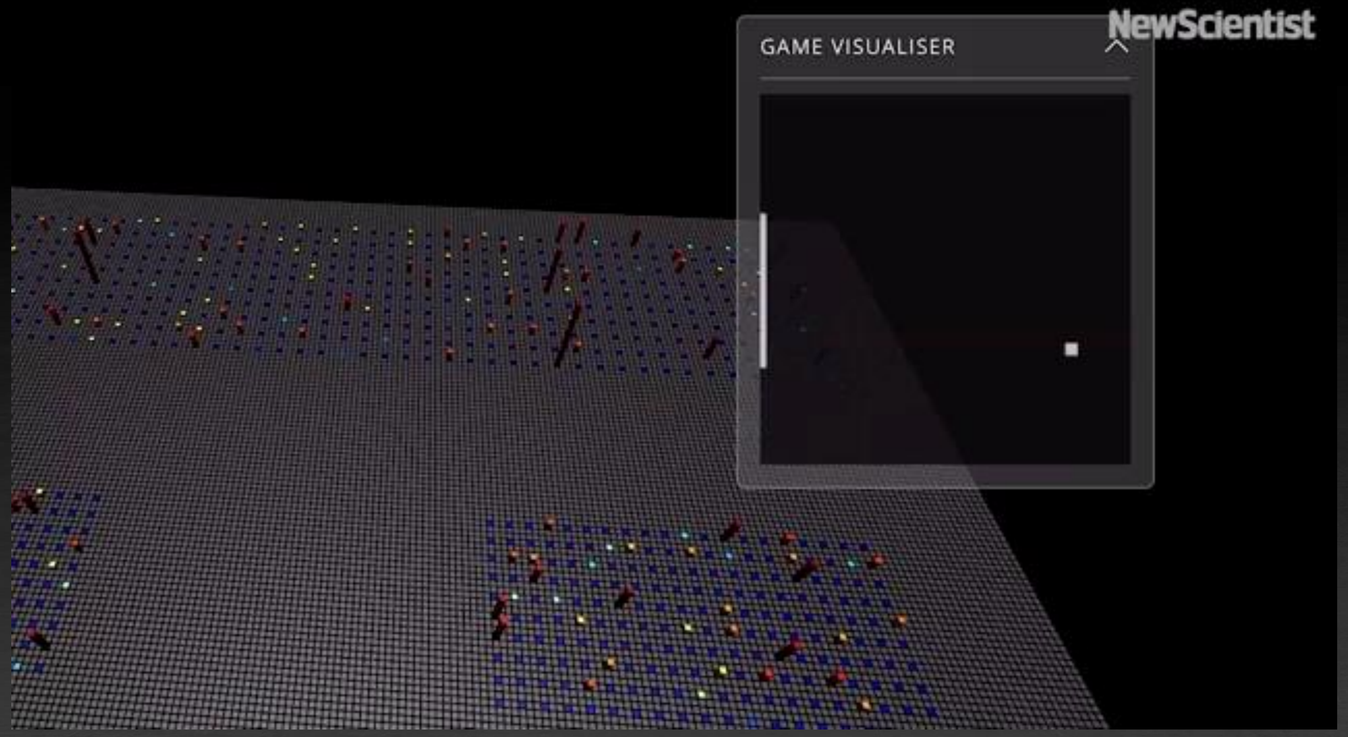
NEWSLETTERS
Sign up to read our regular email newsletters

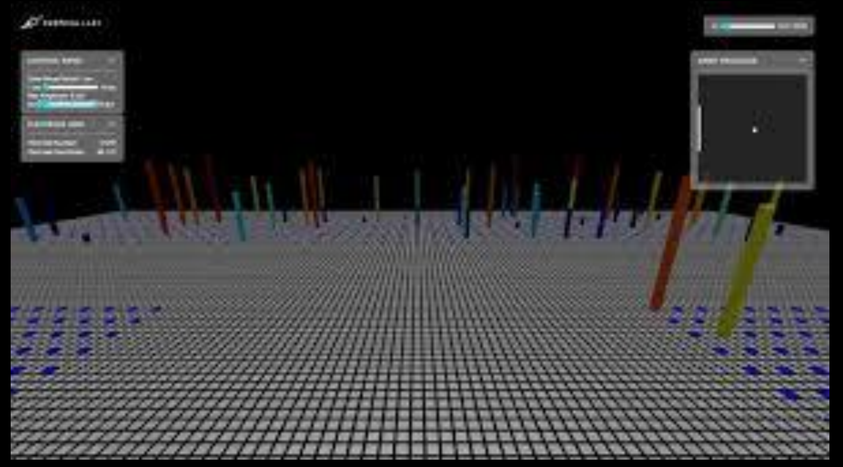
NewScientist

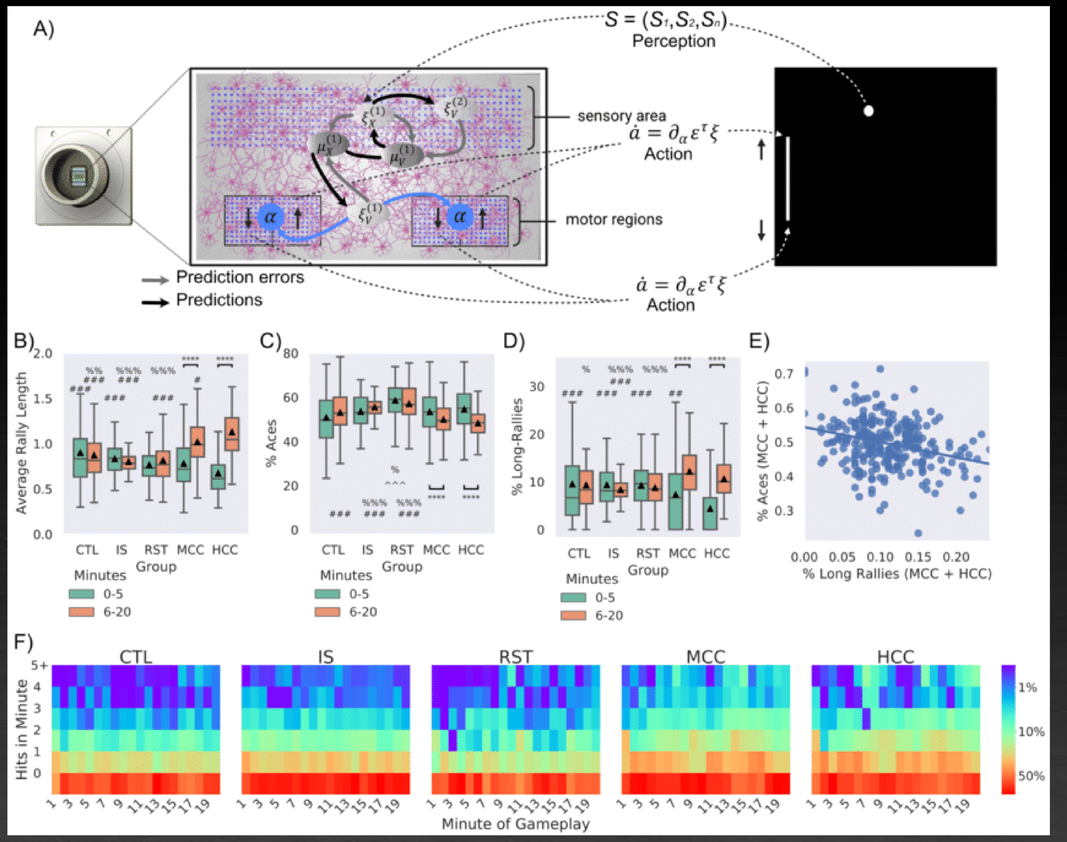
News Podcasts Video Technology Space Physics Health More Shop Courses Events

Human brain cells in a dish learn to play Pong faster than an AI

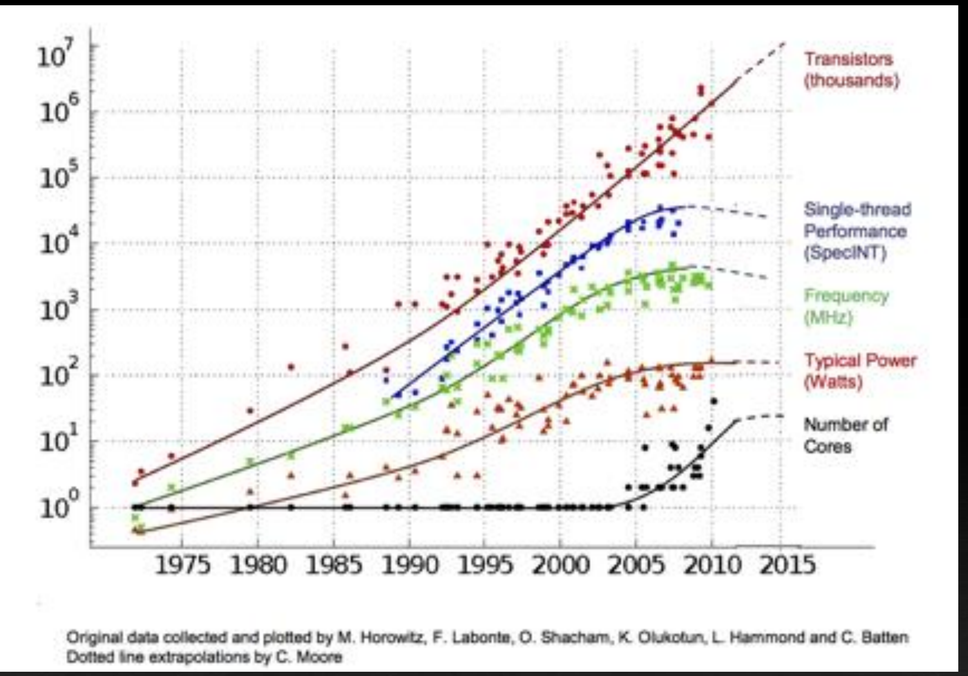
Hundreds of thousands of brain cells in a dish are being taught to play *Pong* by responding to pulses of electricity – and can improve their performance more quickly than an AI can

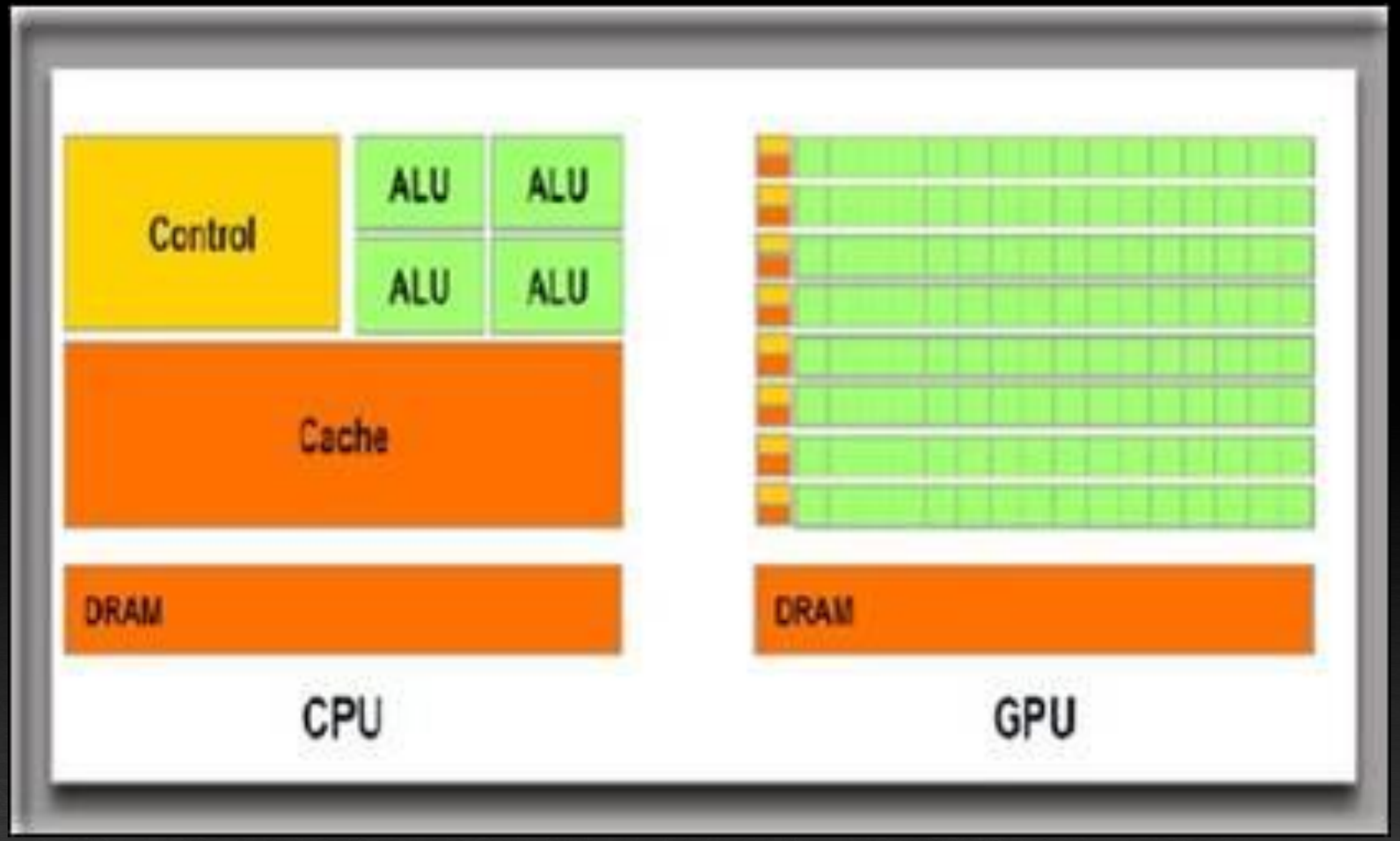




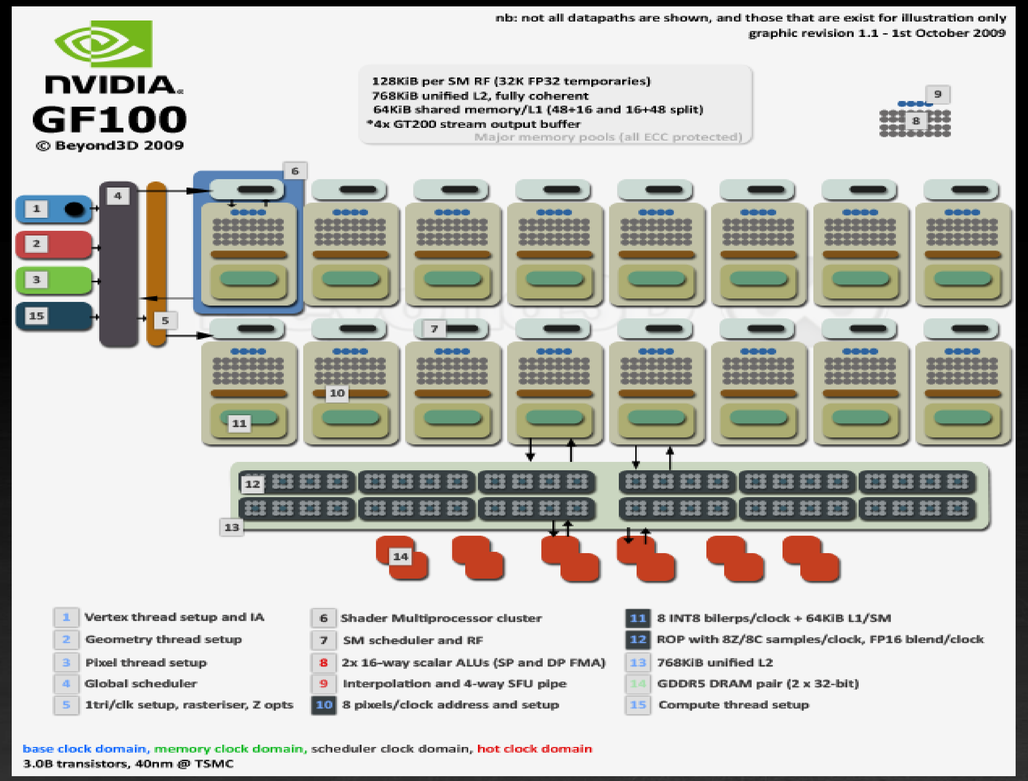


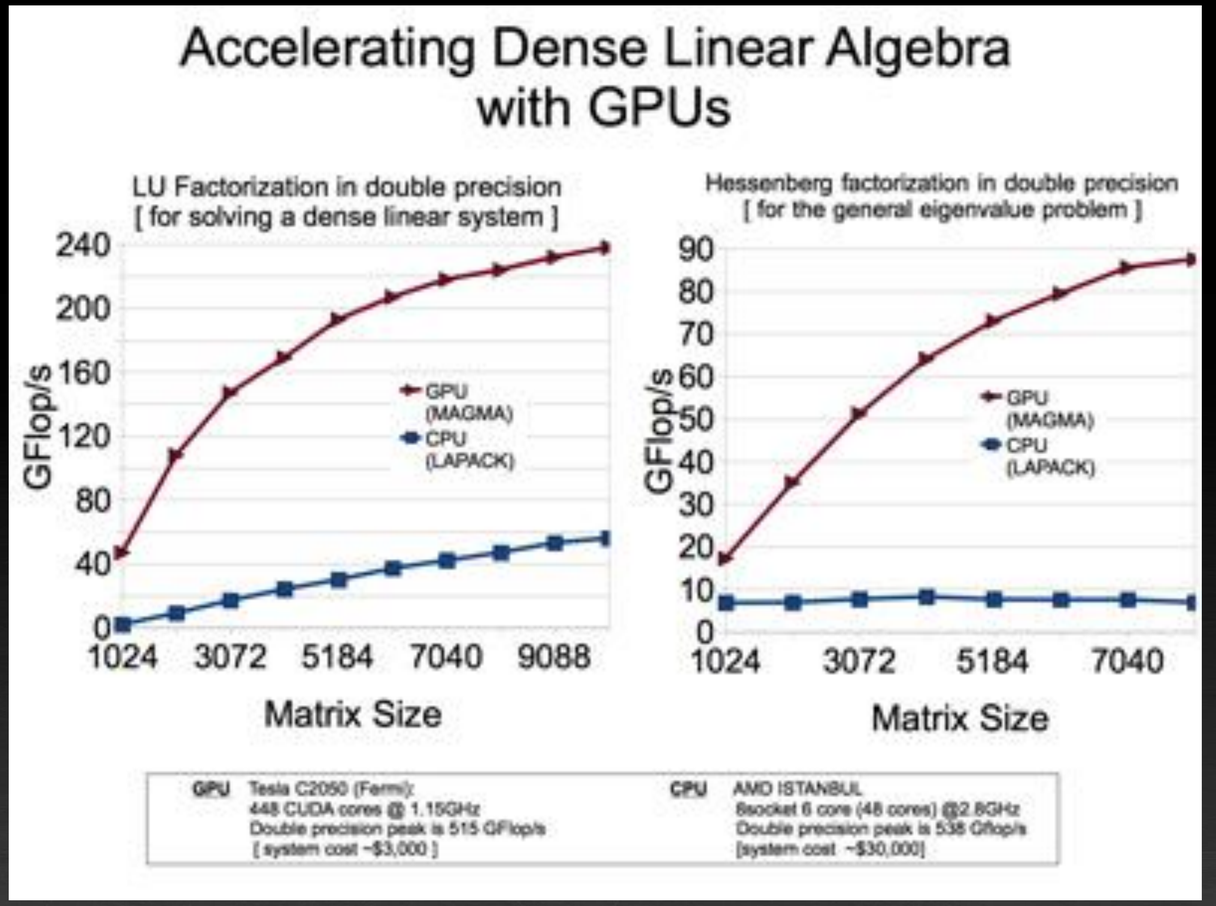
ARCHITECTURAL DIFFERENCE



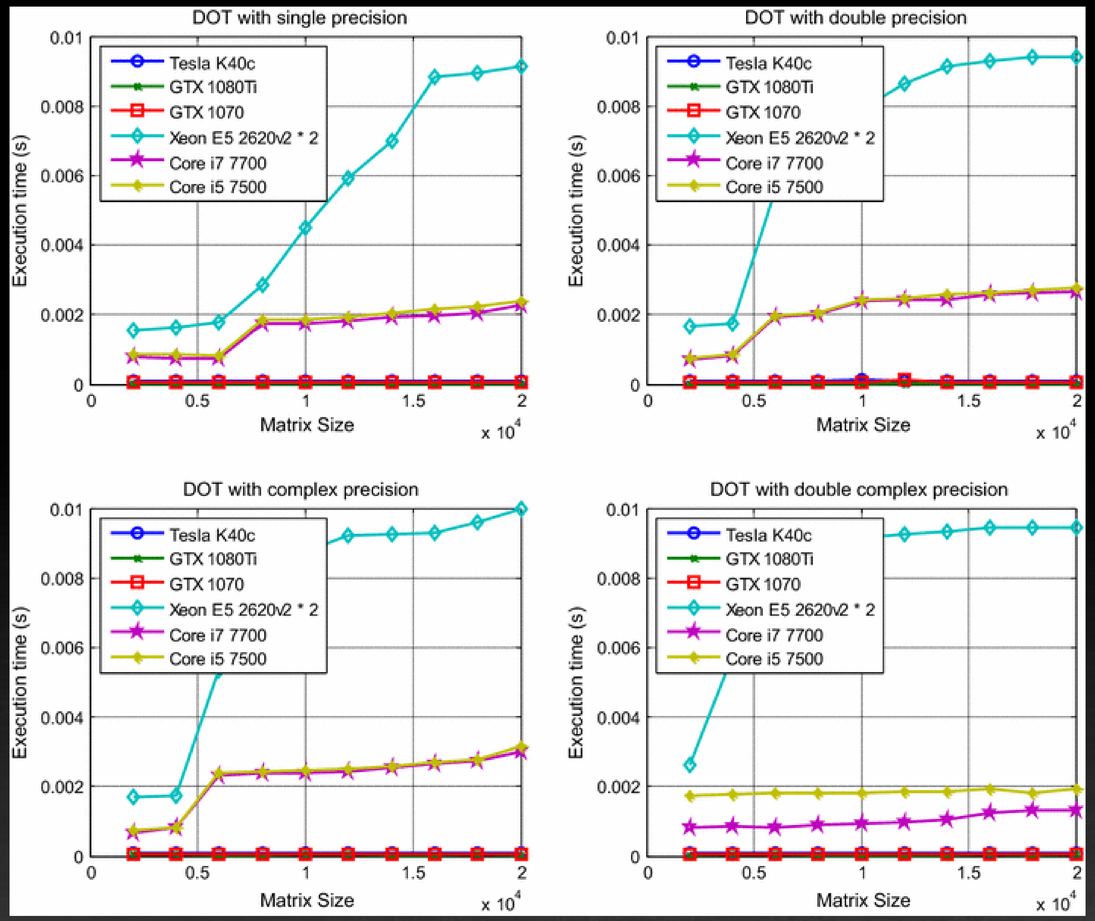


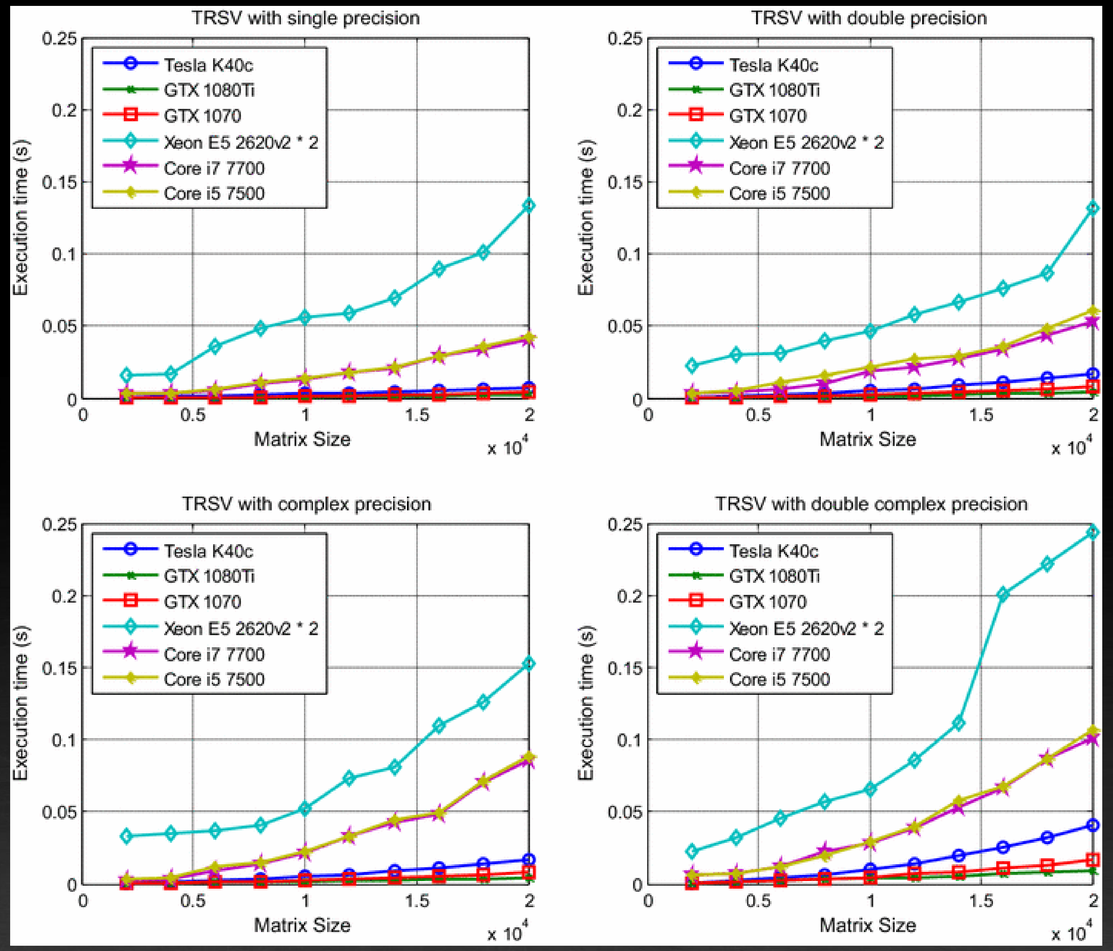
FERMI ARCHITECTURE





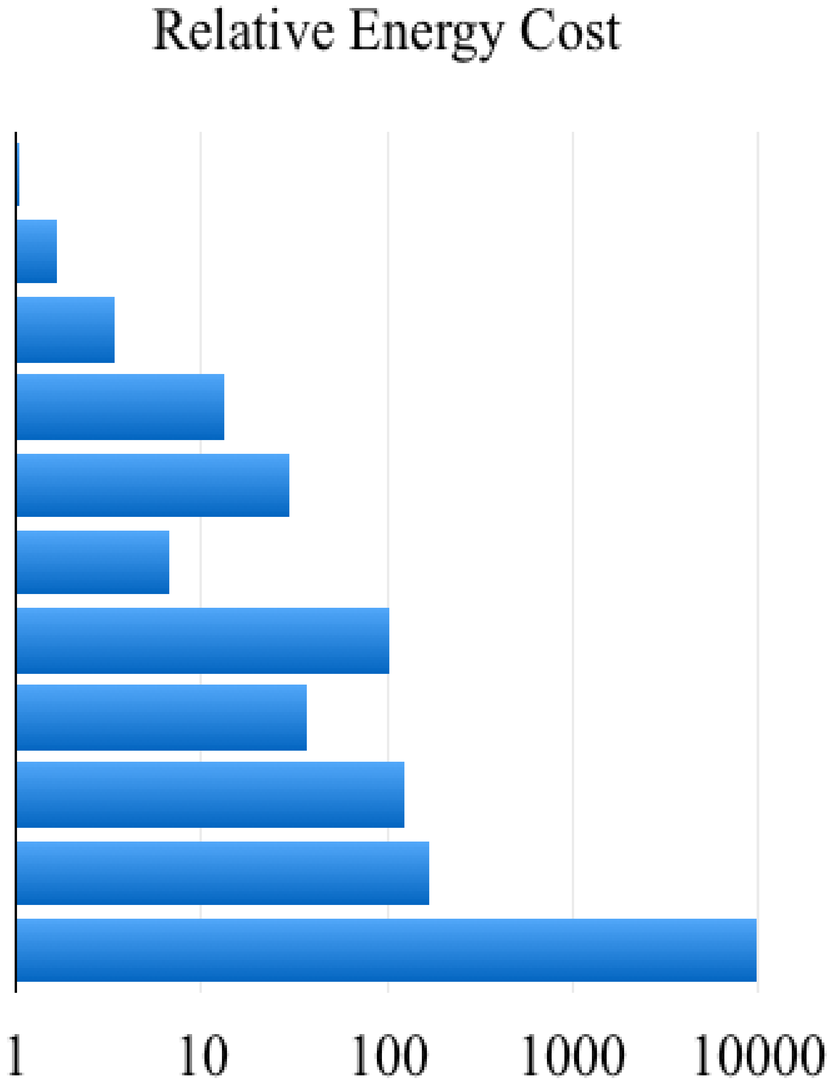
PROOF OF MASSIVELY PARALLEL PROCESSORS

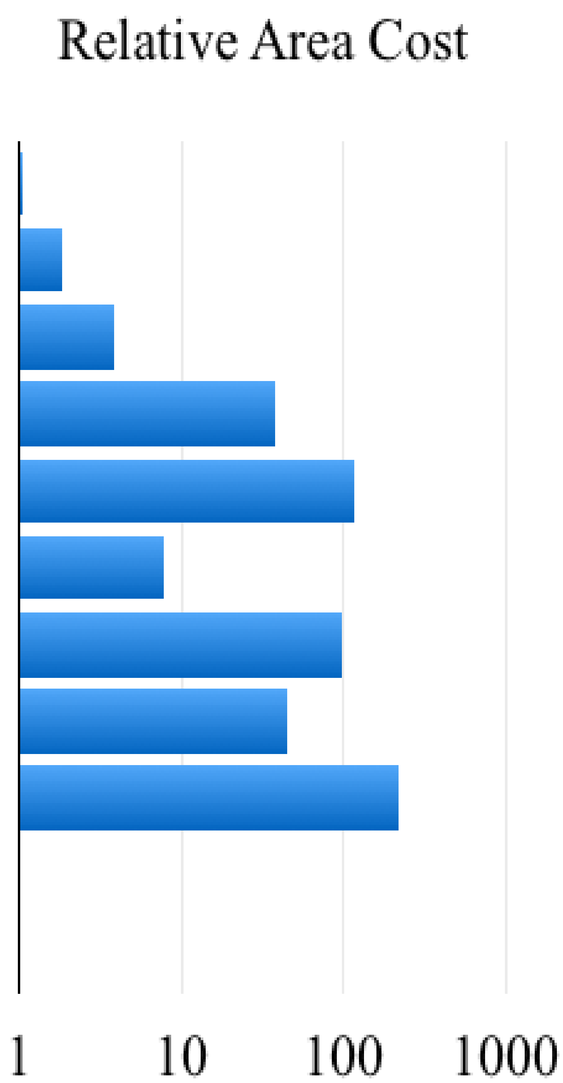




COST OF OPERATIONS

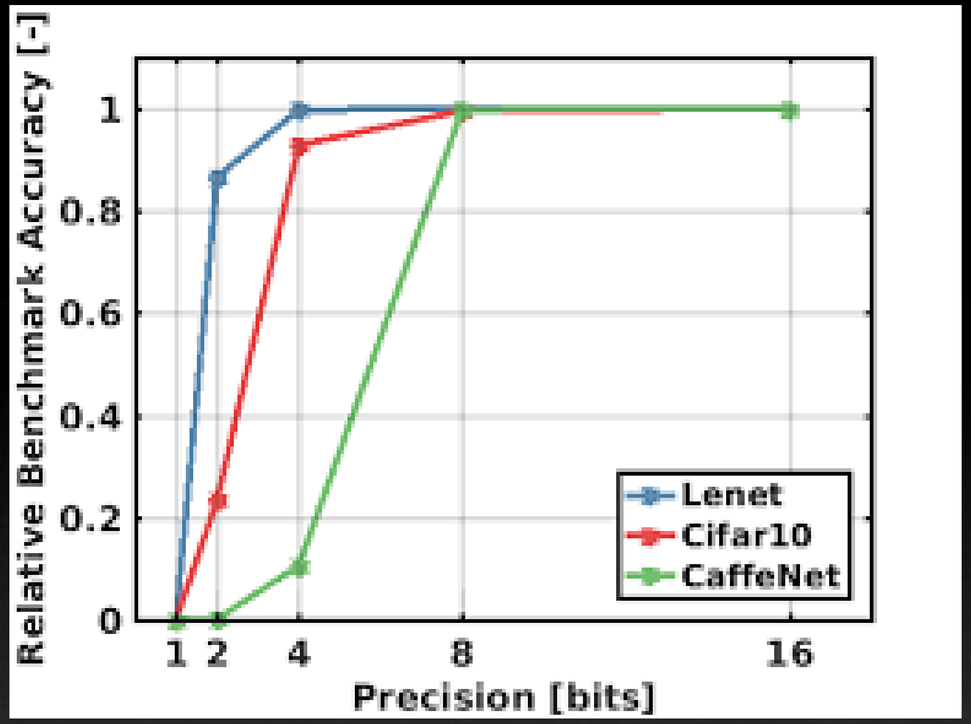
Operation:	Energy (pJ)	Area (μm^2)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A

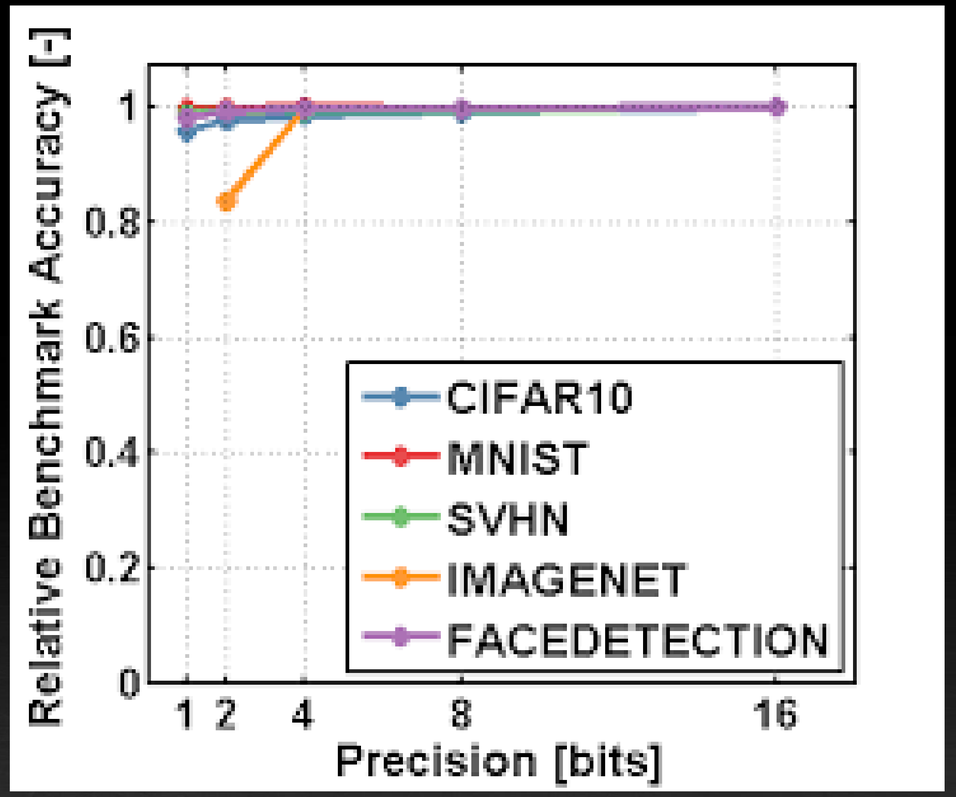




Energy numbers are from Mark Horowitz “Computing’s Energy Problem (and what we can do about it)”, ISSCC 2014
Area numbers are from synthesized result using Design Compiler under TSMC 45nm tech node. FP units used DesignWare Library.

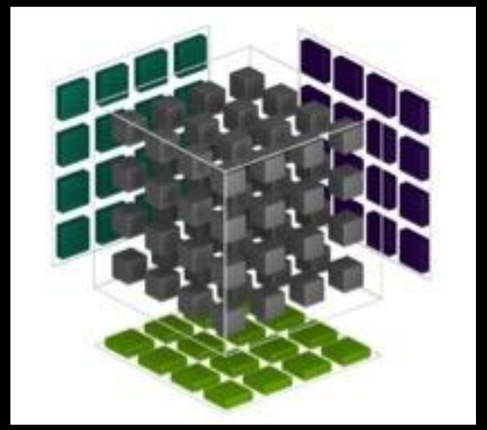
DEEP LEARNING - LESS PRECISION?





TENSOR CORE

Mixed Precision Matrix Math
4x4 matrices

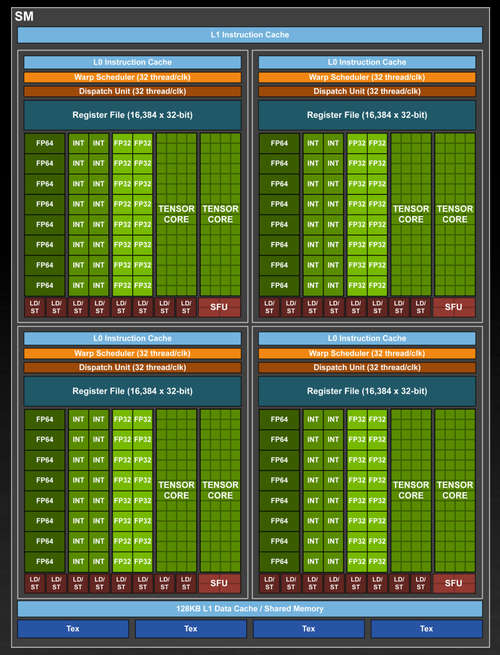


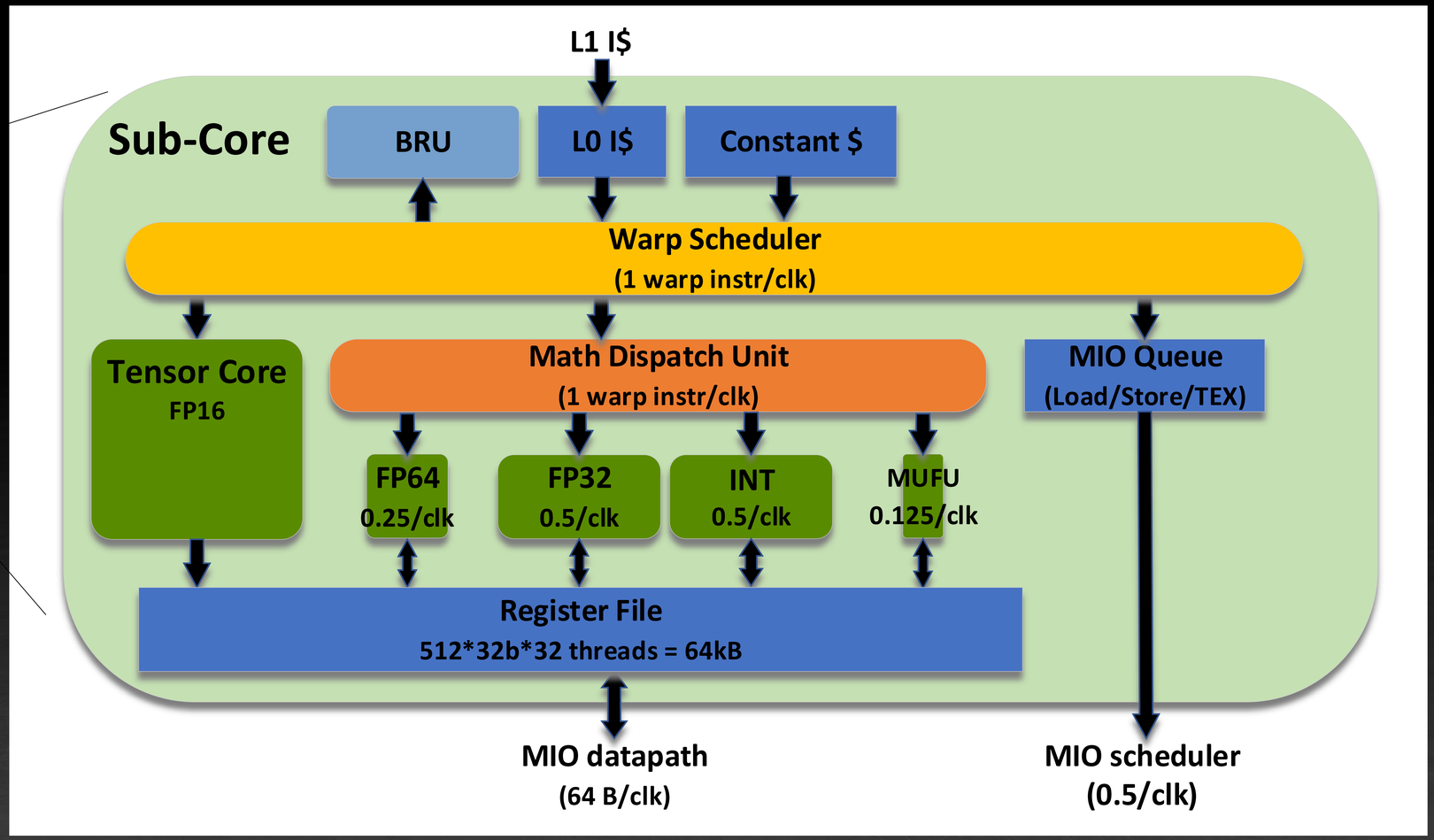
$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

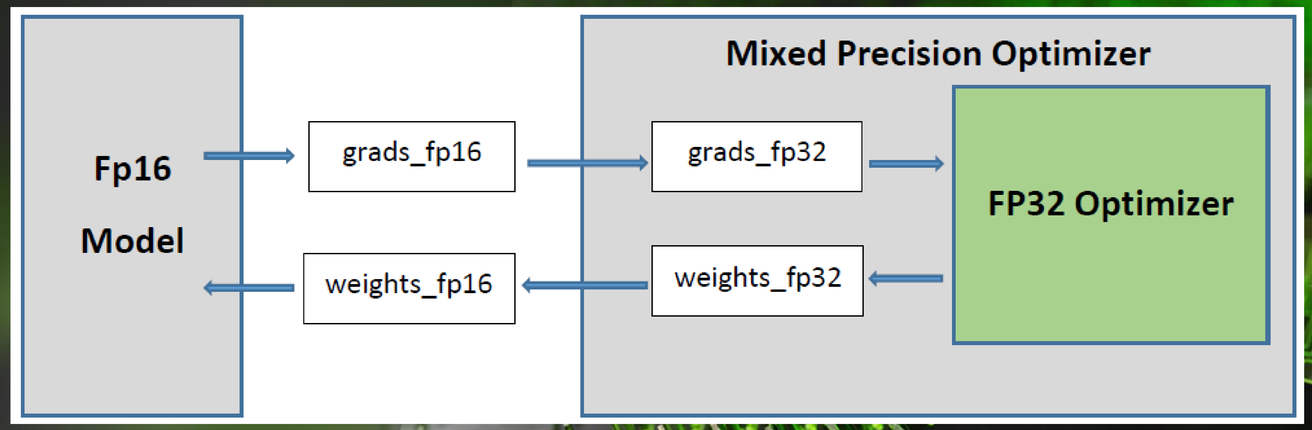
FP16 FP16 FP16 or FP32

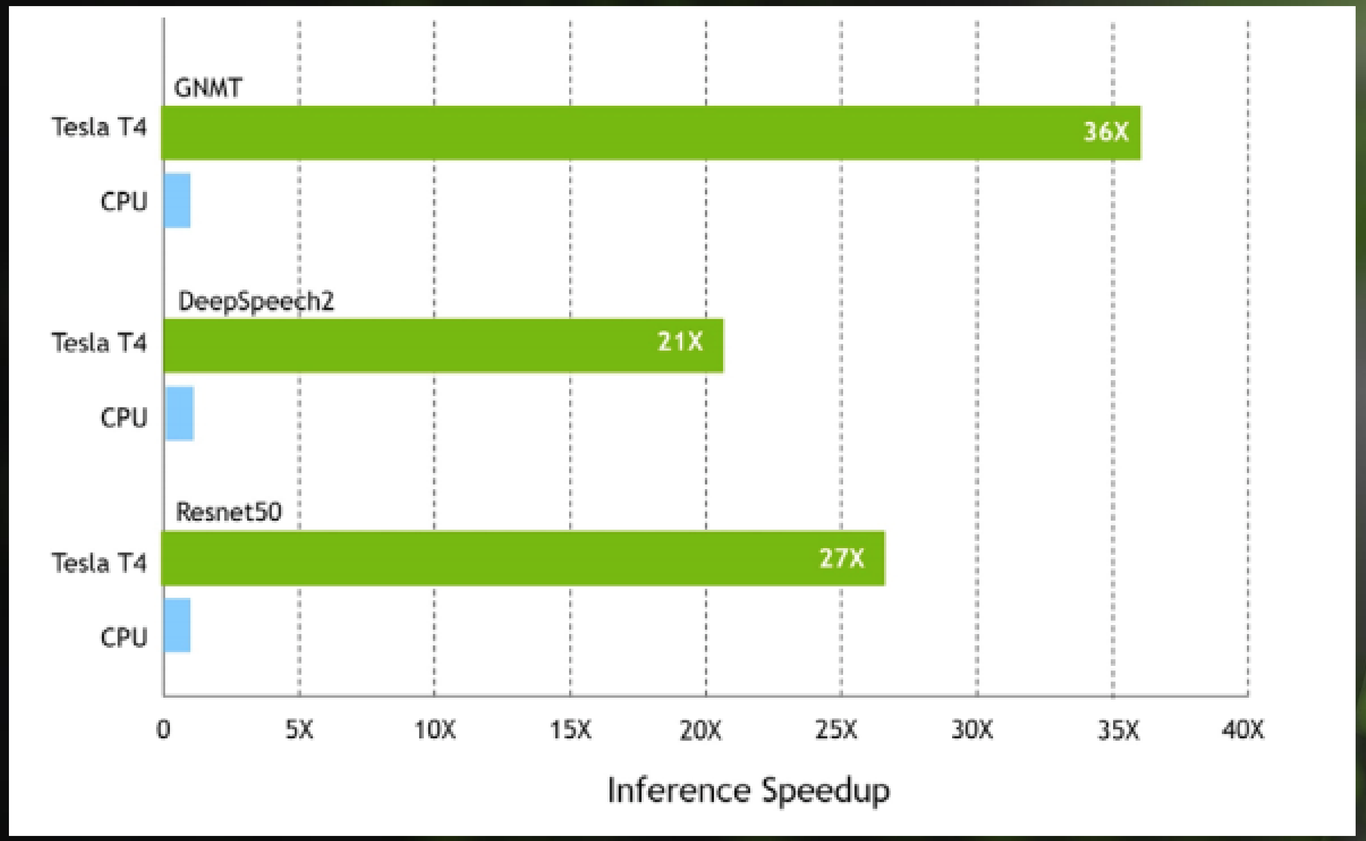
$$D = AB + C$$

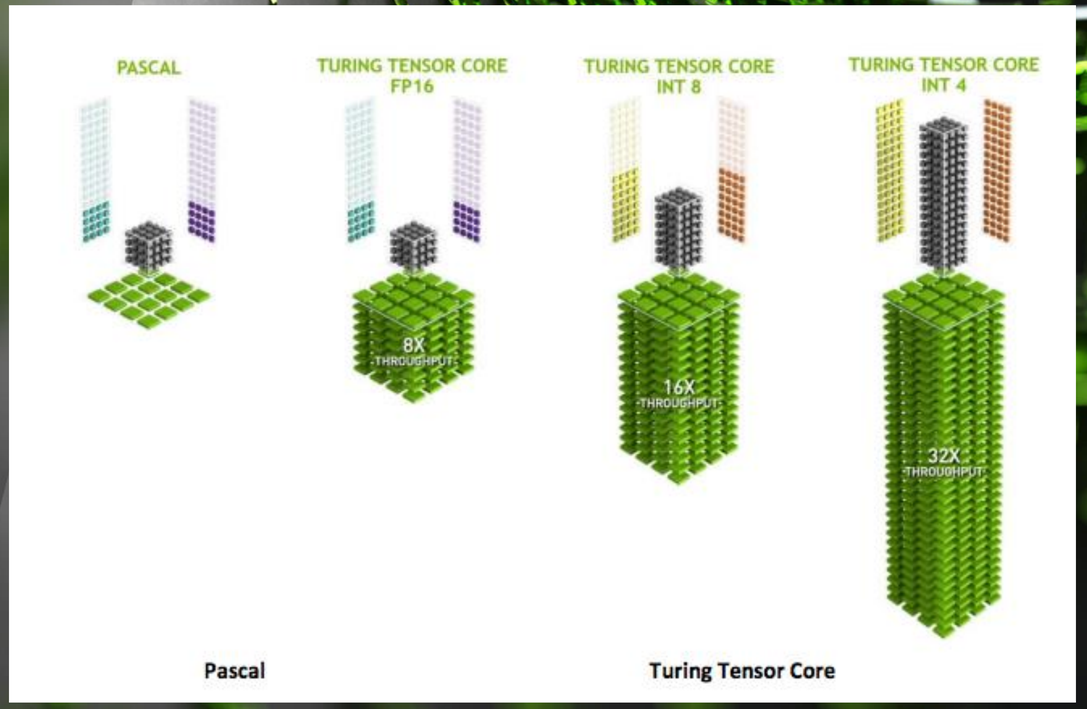
VOLTA SM

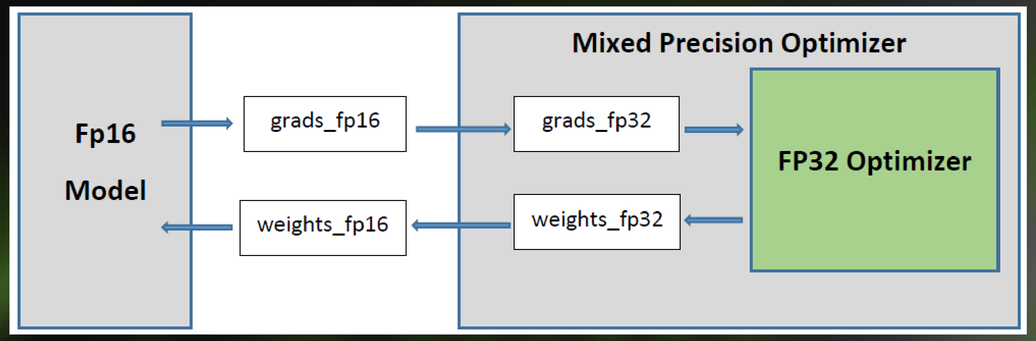

SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex

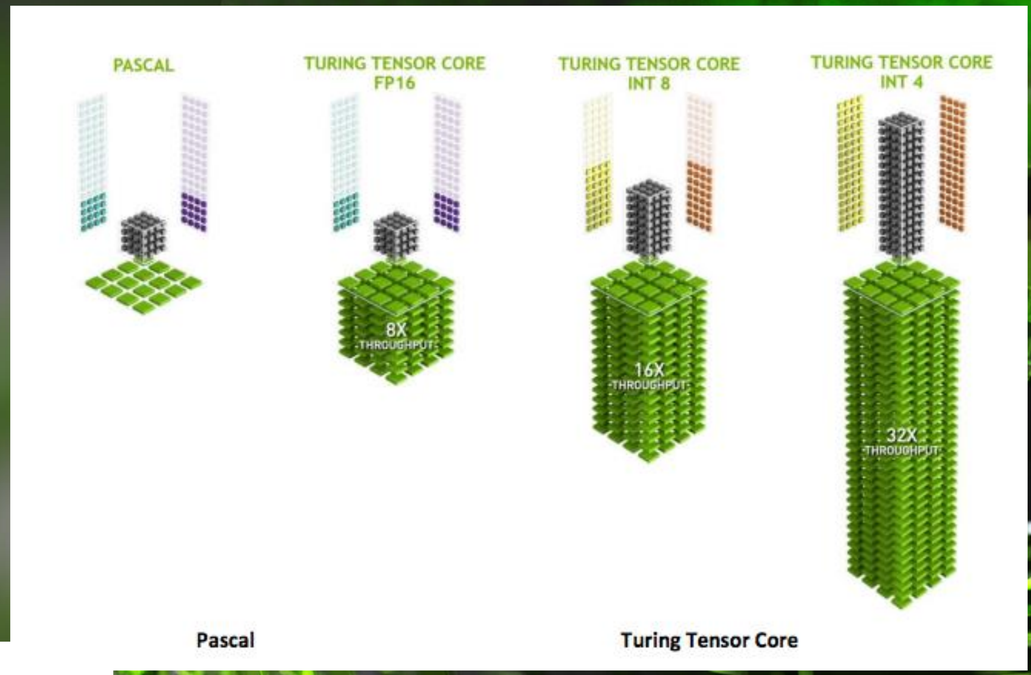

L1 I$
Sub-Core
BRU
L0 I$
Constant $
Warp Scheduler
(1 warp instr/clk)
Tensor Core
FP16
Math Dispatch Unit
(1 warp instr/clk)
MIO Queue
(Load/Store/TEX)
FP64
0.25/clk
FP32
0.5/clk
INT
0.5/clk
MUFU
0.125/clk
Register File
512*32b*32 threads = 64kB
MIO datapath
(64 B/clk)
MIO scheduler
(0.5/clk)

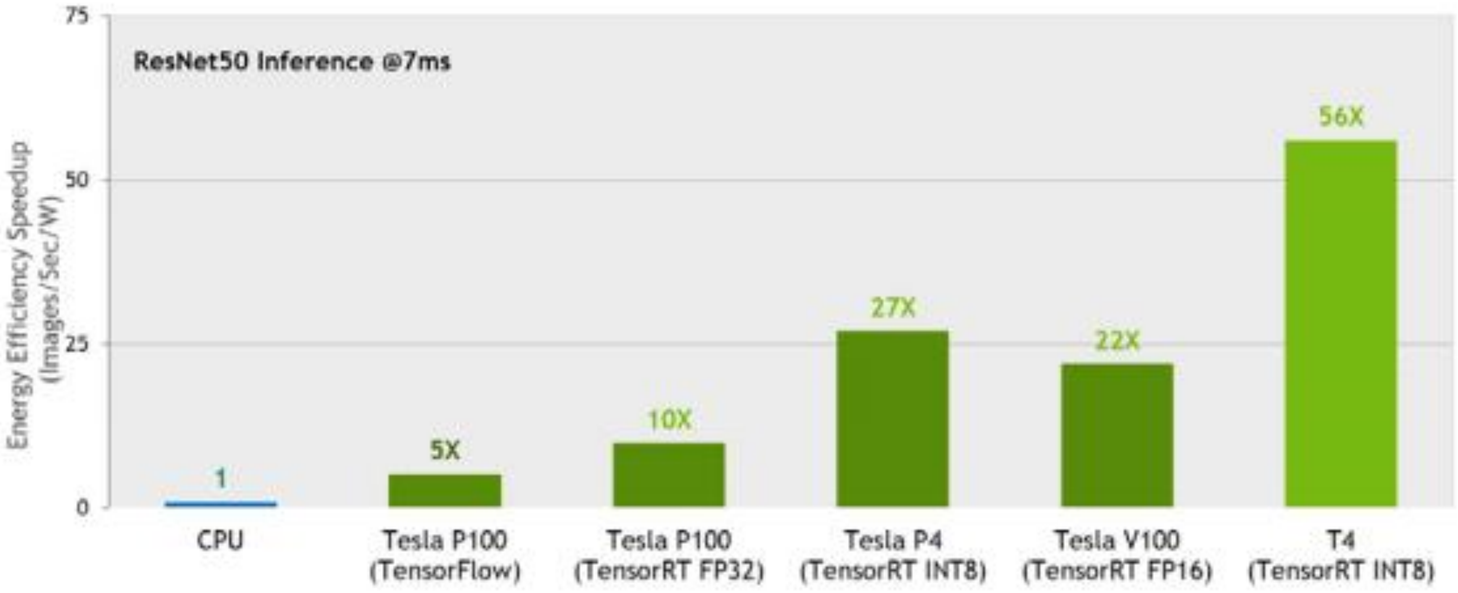

Fp16
Model
grads_fp16
weights_fp16
Mixed Precision Optimizer
grads_fp32
weights_fp32
FP32 Optimizer


GNMT
Tesla T4
36X
CPU
DeepSpeech2
Tesla T4
21X
CPU
Resnet50
Tesla T4
27X
CPU
0
5X
10X
15X
20X
25X
30X
35X
40X
Inference Speedup


PASCAL
TURING TENSOR CORE FP16
TURING TENSOR CORE INT 8
TURING TENSOR CORE INT 4
8X THROUGHPUT
16X THROUGHPUT
32X THROUGHPUT
Pascal
Turing Tensor Core

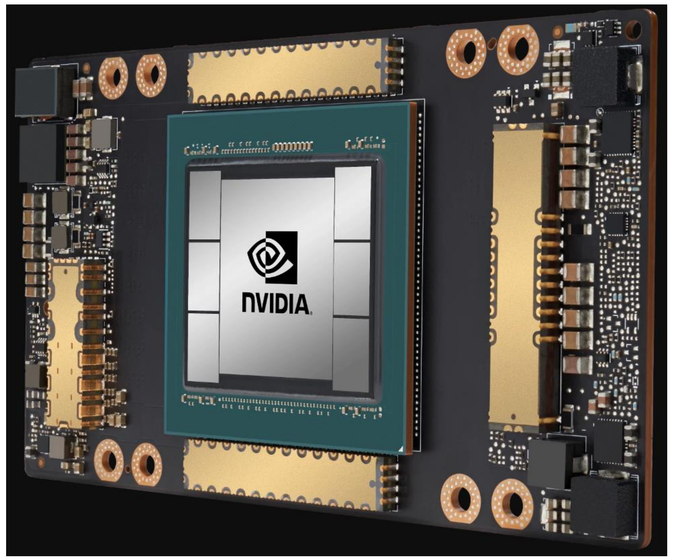

Mixed Precision Optimizer
Fp16
Model
grads_fp16
weights_fp16
grads_fp32
weights_fp32
FP32 Optimizer


PASCAL
TURING TENSOR CORE FP16
TURING TENSOR CORE INT 8
TURING TENSOR CORE INT 4
8X THROUGHPUT
16X THROUGHPUT
32X THROUGHPUT
Pascal
Turing Tensor Core

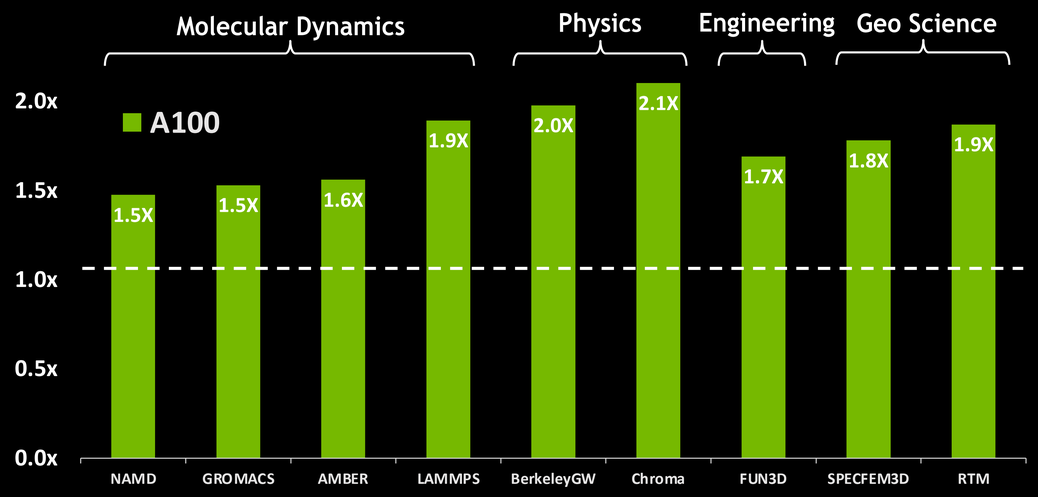

ResNet50 Inference @7ms
Energy Efficiency Speedup (Images/Sec/W)
75
50
25
0
1
5X
10X
27X
22X
56X
CPU
Tesla P100 (TensorFlow)
Tesla P100 (TensorRT FP32)
Tesla P4 (TensorRT INT8)
Tesla V100 (TensorRT FP16)
T4 (TensorRT INT8)

A100 AND V100 SPECS

	V100	A100
SMs	80	108
Tensor Core Precision	FP16	FP64, TF32, BF16, FP16, I8, I4, B1
Shared Memory per Block	96 kB	160 kB
L2 Cache Size	6144 kB	40960 kB
Memory Bandwidth	900 GB/sec	1555 GB/sec
NVLink Interconnect	300 GB/sec	600 GB/sec
FP64 Throughput	7.9 TFLOPS	9.7 | 19.5 TFLOPS
TF32 Tensor Core	N/A	156 | 312 TFLOPS



1.5-2X MORE HPC PERFORMANCE THAN V100

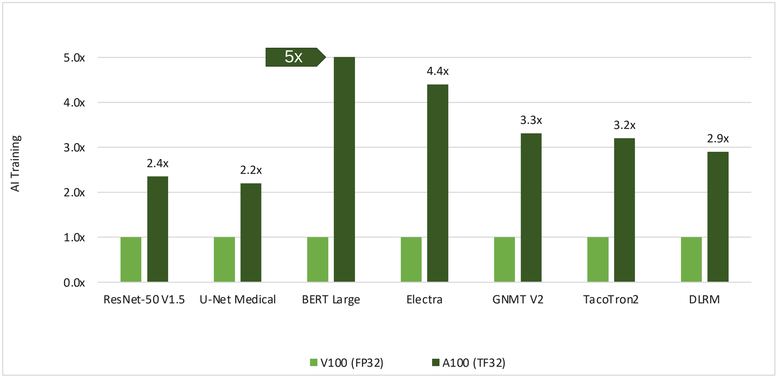


All results are measured
Except BerkeleyGW, *V100 used is single V100 SXM2. A100 used is single A100 SXM4*
More apps detail: AMBER based on PME-Cellulose, GROMACS with STMV (h-bond), LAMMPS with Atomic Fluid LJ-2.5, NAMD with v3.0a1 STMV_NVE
Chroma with szscl21_24_128, FUN3D with dpw, RTM with Isotropic Radius 4 1024^3, SPECFEM3D with Cartesian four material model
BerkeleyGW based on Chi Sum and uses 8xV100 in DGX-1, vs 8xA100 in DGX A100

3RD GENERATION MULTI-PRECISION TENSOR CORES IN A100

Greatest Generational Leap – 20X Volta

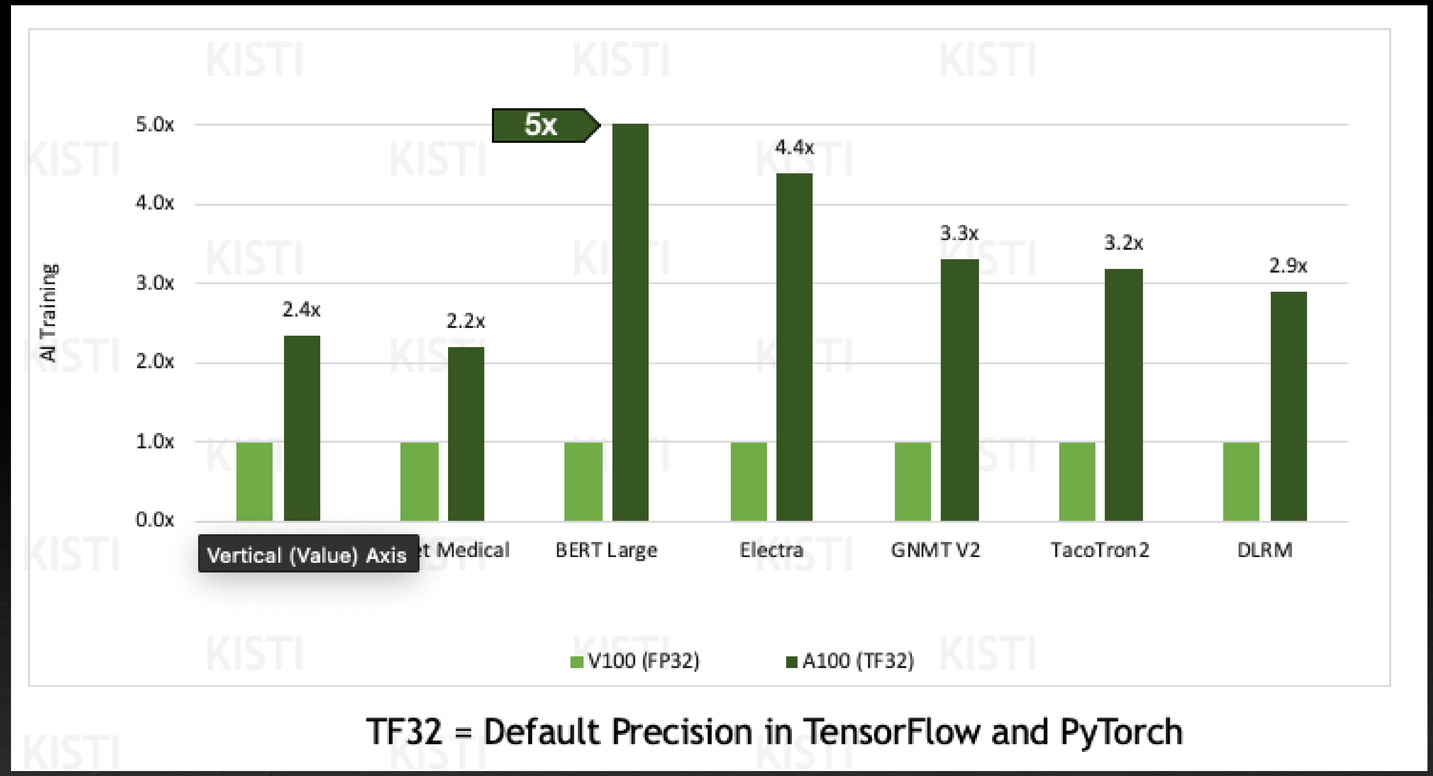
Precision	Peak	Vs Volta
FP64 HPC	19.5 TFLOPS	2.5x
FP32 TRAINING	312 TFLOPS	20x
INT8 INFERENCE	1,248 TOPS	20x



TF32 = Default Precision in TensorFlow and PyTorch

Greatest Generational Leap – 20X Volta

Precision	Peak	Vs Volta
FP64 HPC	19.5 TFLOPS	2.5x
FP32 TRAINING	312 TFLOPS	20x
INT8 INFERENCE	1,248 TOPS	20x



TF32 = Default Precision in TensorFlow and PyTorch


SCIENTIFIC CHALLENGES

SCIENTIFIC CHALLENGES

Problems that arise when applying AI to science

- Data Labelling
- Limited Data
- Enforcing Physical Constraints
- Uncertainty Quantification
- Trustworthiness and Interpretability
- HPC - AI Coupling
- Loss of Dynamic Range
- Data Movement
- Numerical Stability
- Generalization



MILLION-X CLIMATE SCIENCE

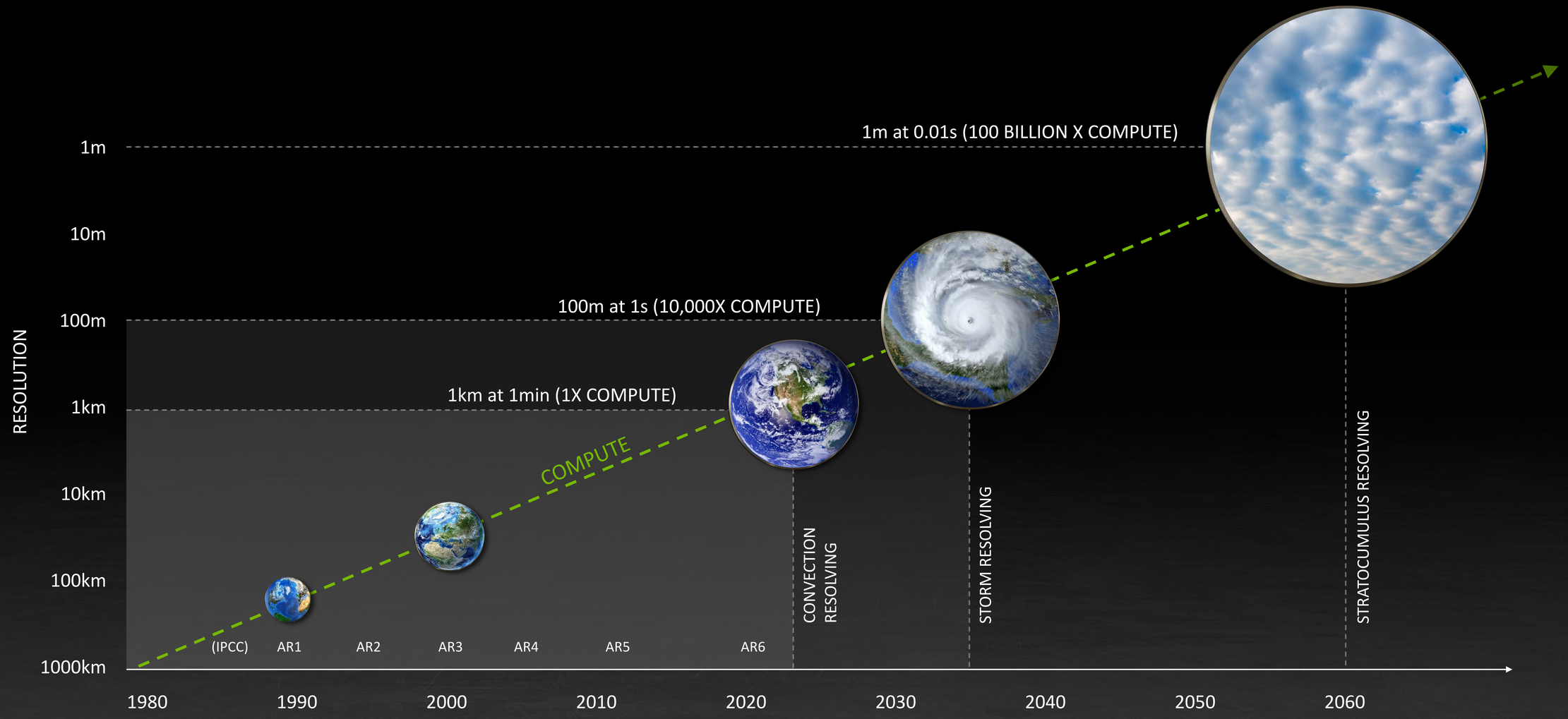


Figure adapted from: Schneider, T., Teixeira, J., Bretherton, C. *et al.* Climate goals and computing the future of clouds. *Nature Clim Change* **7,** 3–5 (2017). https://doi.org/10.1038/nclimate3190

NVIDIA MODULUS

Physics-ML Neural Simulation Framework

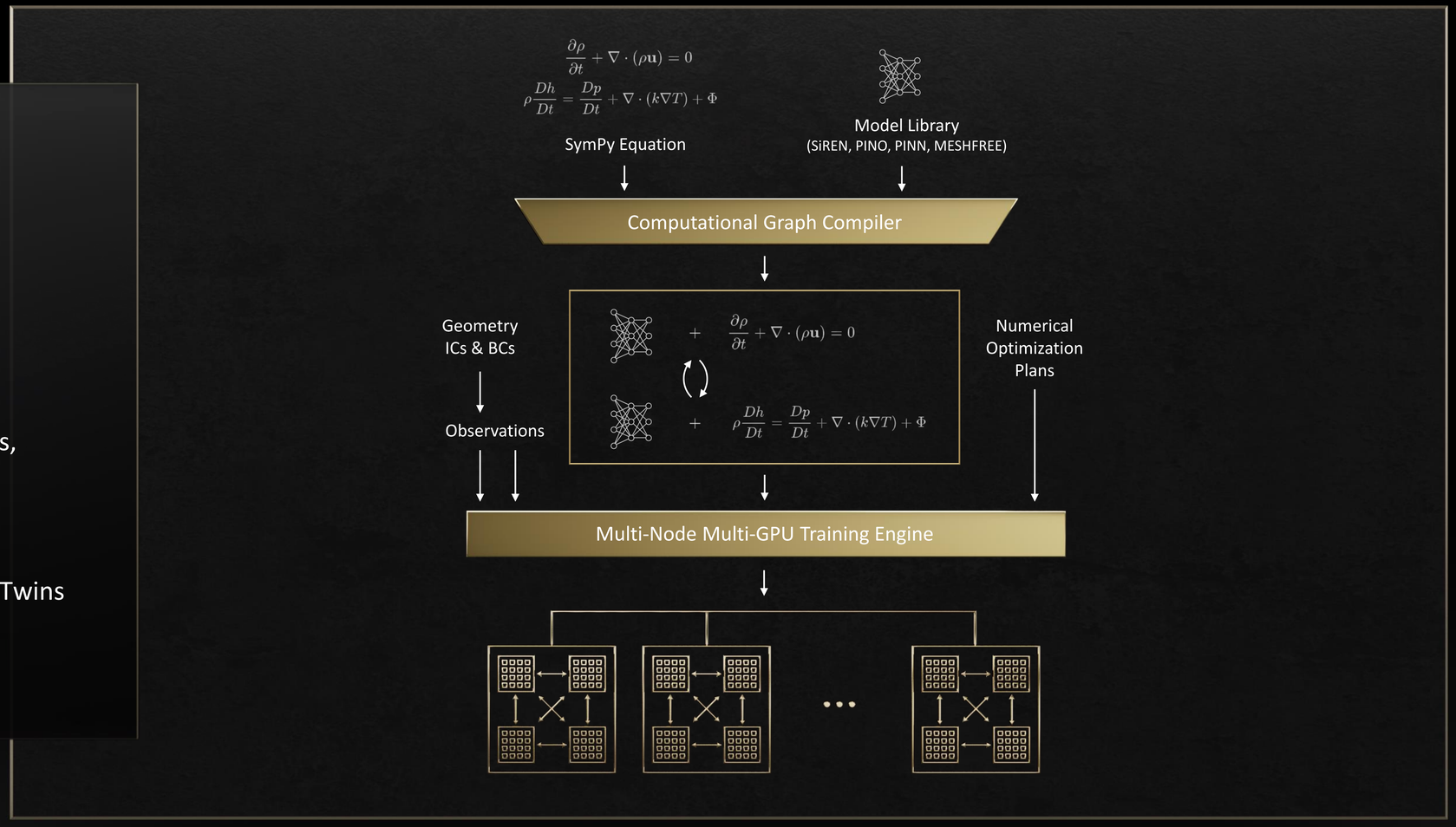
Framework for Developing Physics-ML Models

Train Physics-ML Models Using Governing Physics, Simulation, and Observed Data

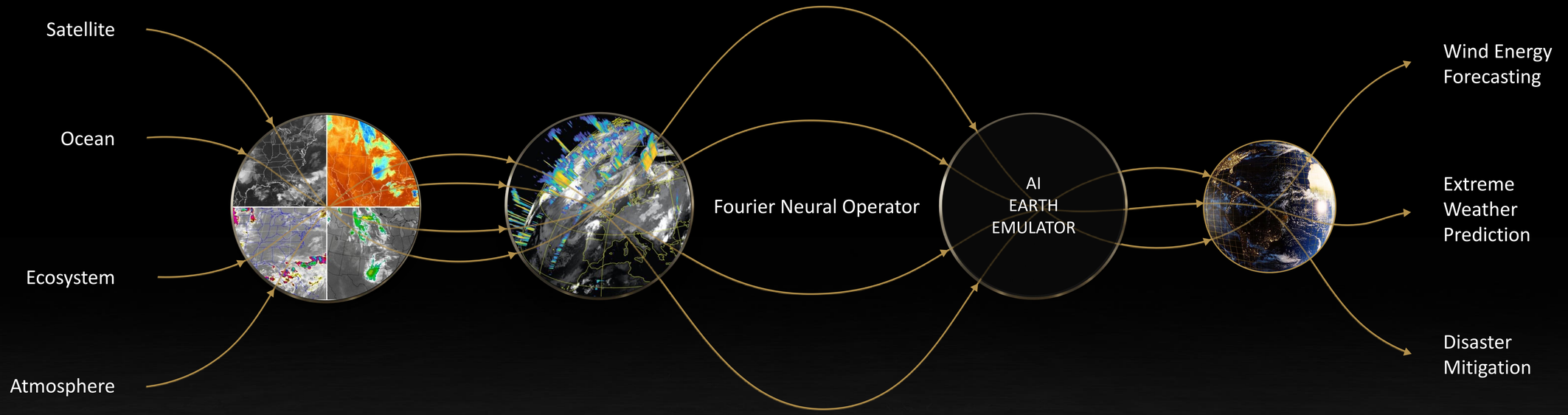
Multi-GPU, Multi-Node Training

1,000-100,000X Speed Models – Ideal for Digital Twins

Available Now
developer.nvidia.com/modulus



EARTH DIGITAL TWIN IN OMNIVERSE



ERA5 ECMWF
Atmospheric Winds & Geopotential
10 TB | 30km | 5 Atmos Layers

100,000X Speed-Up
0.25 Seconds for 7-Day Forecast
Training: 4 Hours on 128 A100 GPUs

RAPIDS — Modulus — Omniverse

Visualizing a Twin Earth

in NVIDIA Omniverse

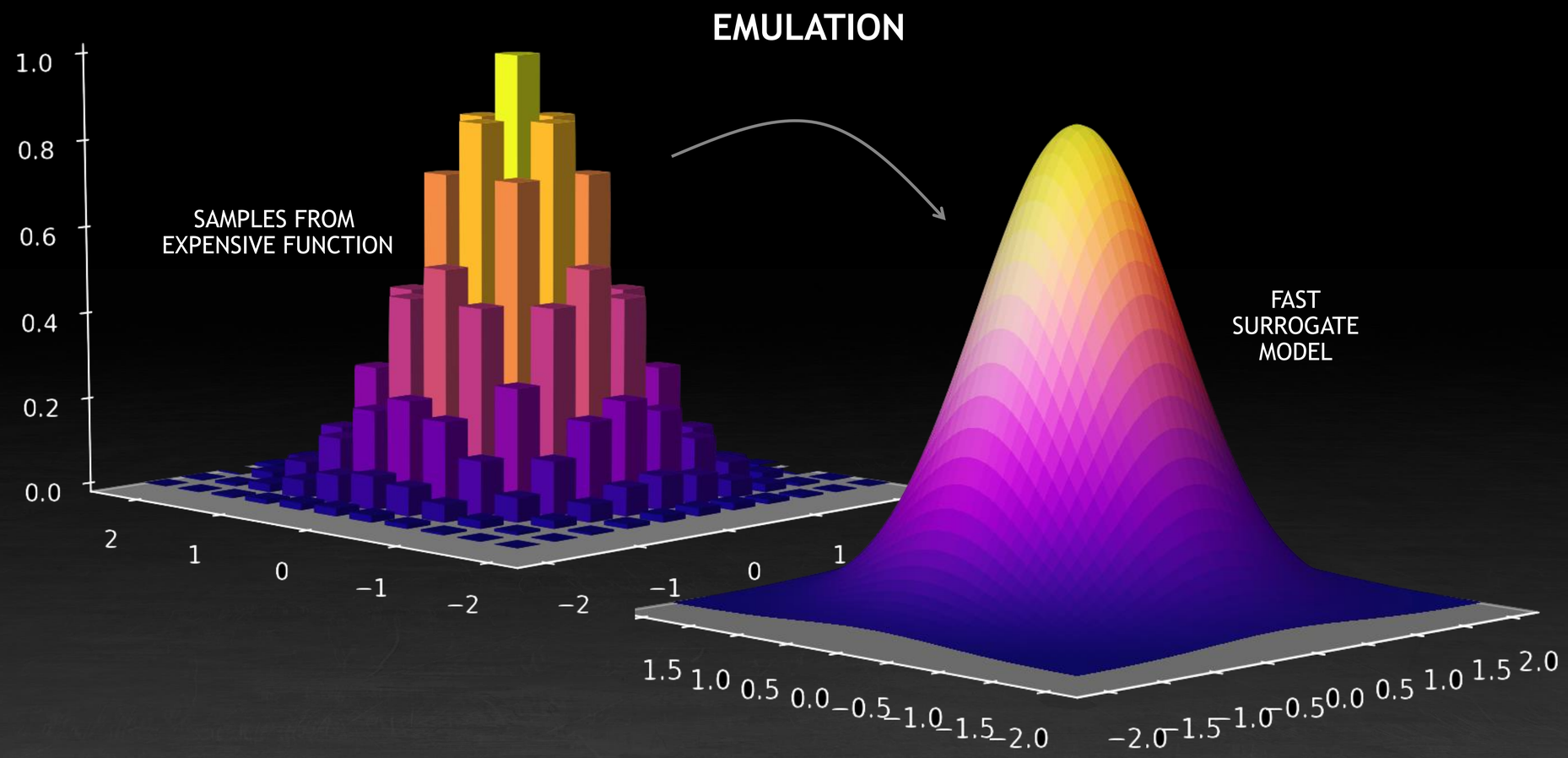




NVIDIA®

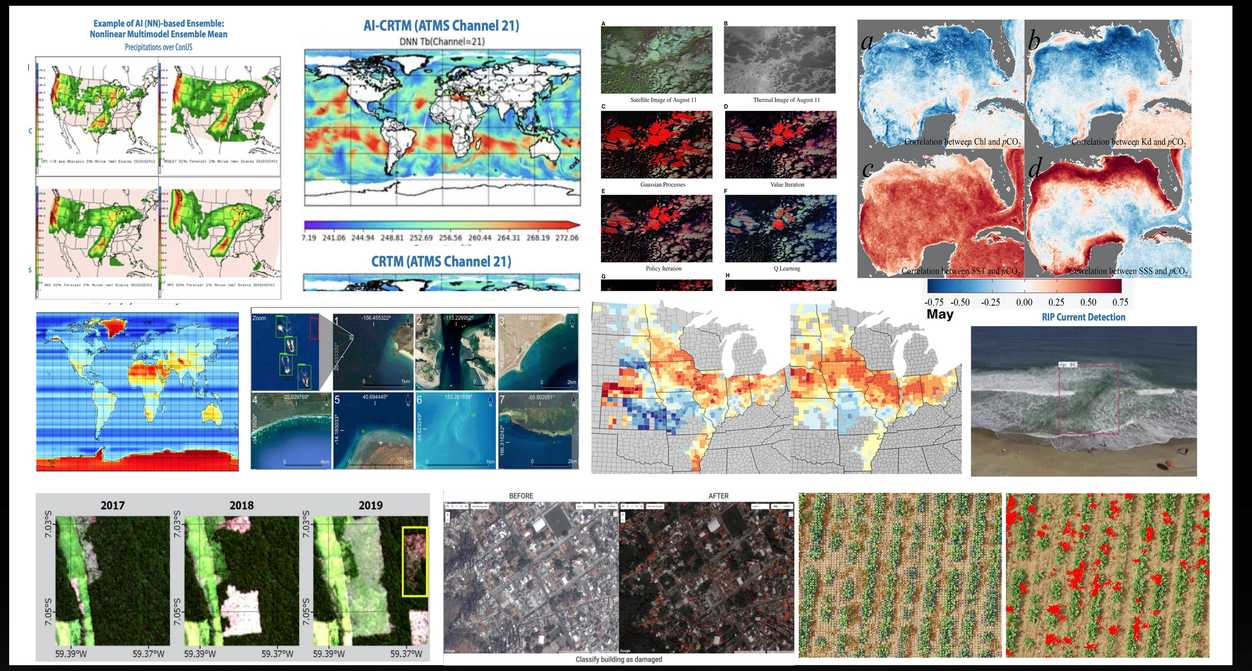
ACCELERATED PHYSICS

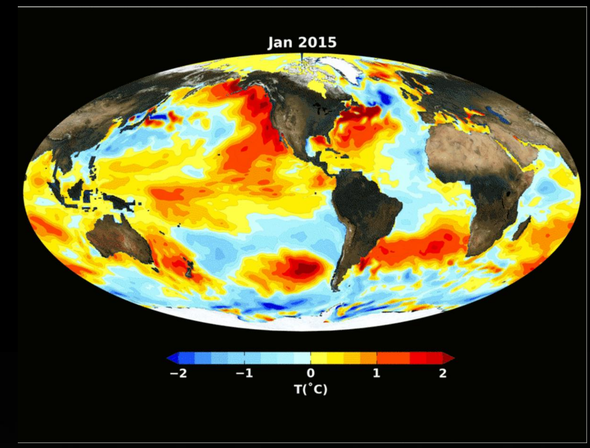
Using surrogate models to speed up existing code

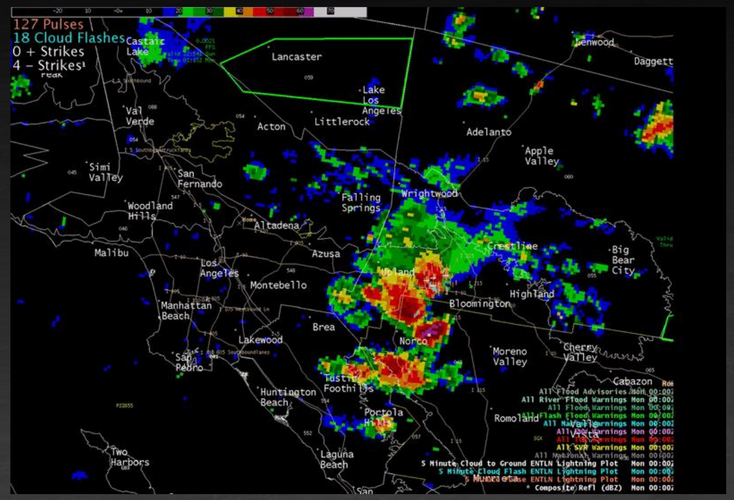


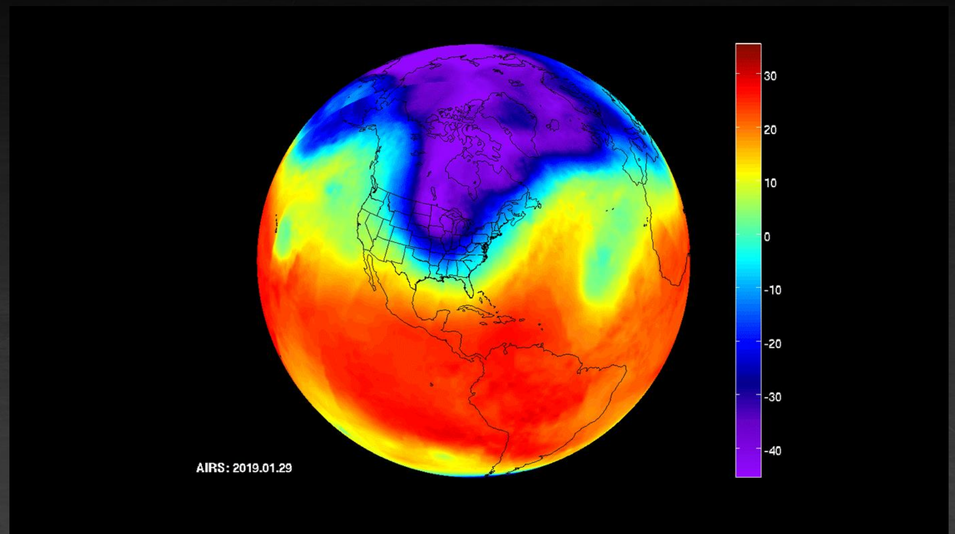
Deep learning is being rapidly adopted by the Earth System Science community













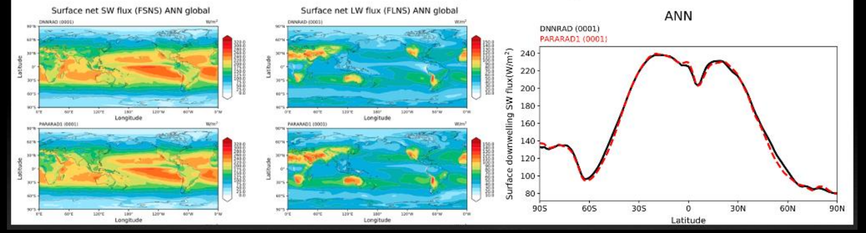
ACCELERATED PHYSICS PARAMETERIZATIONS

Emulation of E3SM Super-parameterized SW and LW Radiation

Using Deep Neural Networks as Cost-Effective Surrogate Models for Super-Parameterized E3SM Radiative Transfer

Anikesh Pal[1], Salil Mahajan[2], and Matthew R. Norman[1]

[1]National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA, [2]Computational Earth SciencesOak Ridge National Laboratory, Oak Ridge, TN, USA

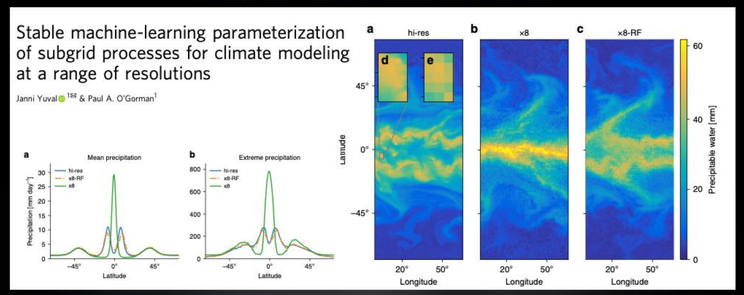


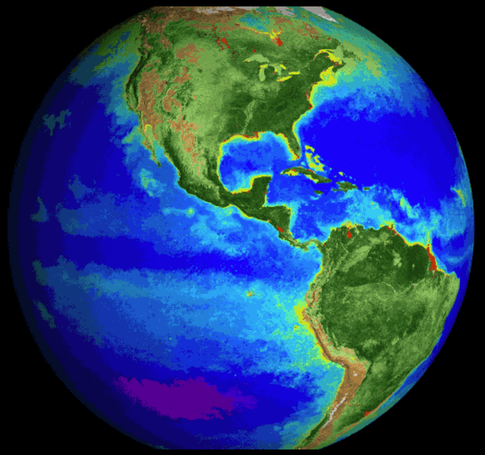
IMPROVED PHYSICS PARAMETERIZATIONS

Building more accurate physics parameterizations

Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions

Janni Yuval[1&8] & Paul A. O'Gorman[1]

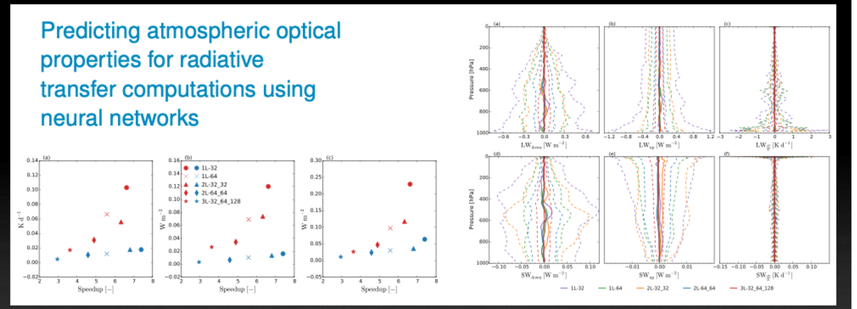




ACCELERATED PHYSICS PARAMETERIZATIONS

Emulation of the RRTMGP radiative transfer model

Predicting atmospheric optical properties for radiative transfer computations using neural networks



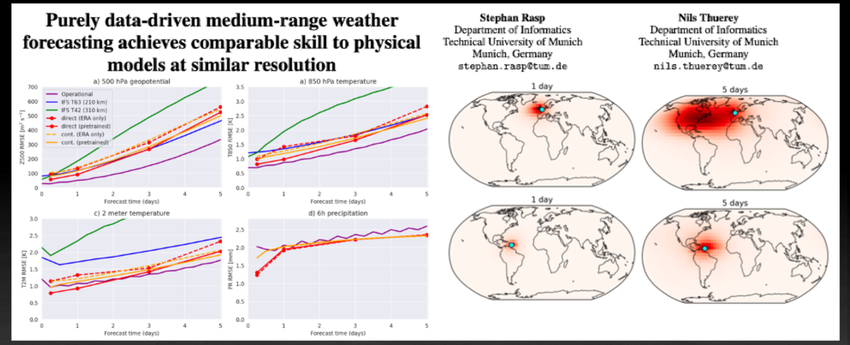
MEDIUM-RANGE WEATHER FORCEGASTING

Data Driven models match accuracy of physical models on weather-bench

Purely data-driven medium-range weather forecasting achieves comparable skill to physical models at similar resolution

Stephan Rasp
Department of Informatics
Technical University of Munich
Munich, Germany
stephan.rasp@tum.de

Nils Thuerey
Department of Informatics
Technical University of Munich
Munich, Germany
nils.thuerey@tum.de



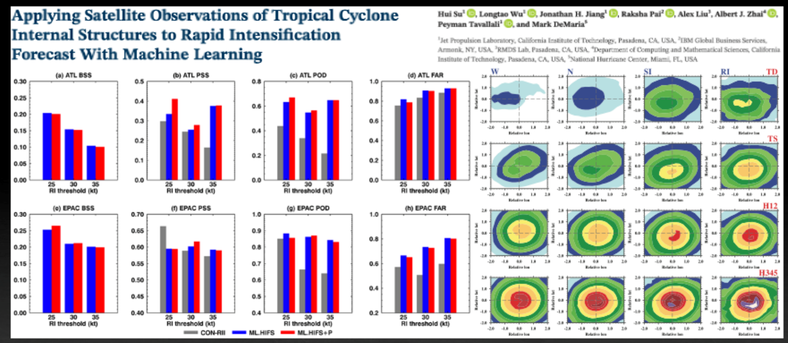
RAPID INTENSIFICATION

Machine Learning improves detection by 40-200 % over operational consensus

Applying Satellite Observations of Tropical Cyclone Internal Structures to Rapid Intensification Forecast With Machine Learning

Hui Su[1], Longtao Wu[1], Jonathan H. Jiang[1], Rakshu Pai[2], Alex Liu[1], Albert J. Zhai[2], Peyman Tavallali[1], and Mark DeMaria[3]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA, [2]IBM Global Business Services, Armonk, NY, USA, [3]RMDI Lab, Pasadena, CA, USA, [4]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA, [5]National Hurricane Center, Miami, FL, USA

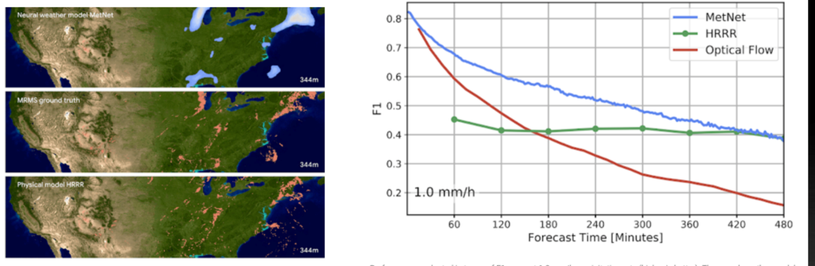


NOWCASTING

MetNet: Can beat physical models up to 8 hours

MetNet: A Neural Weather Model for Precipitation Forecasting

GOOGLE RESEARCH



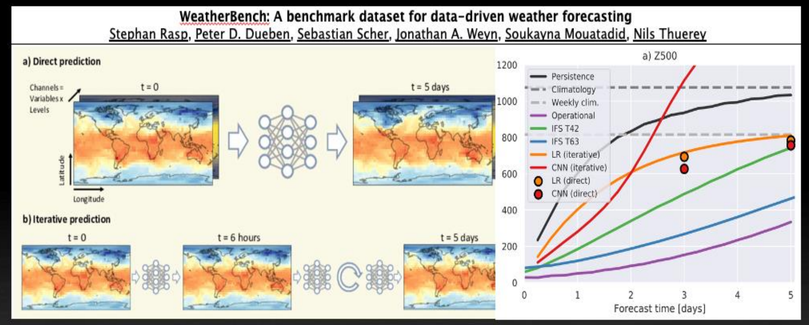
Performance evaluated in terms of F1-score at 1.0 mm/h precipitation rate (higher is better). The neural weather model (MetNet) outperforms the physics-based model (HRRR) currently operational in the US for timescales up to 8 hours ahead.

MEDIUM-RANGE WEATHER FORECASTING

Weather Bench: A Standardized Benchmark for Data Driven Forecasts

WeatherBench: A benchmark dataset for data-driven weather forecasting

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, Nils Thuerey

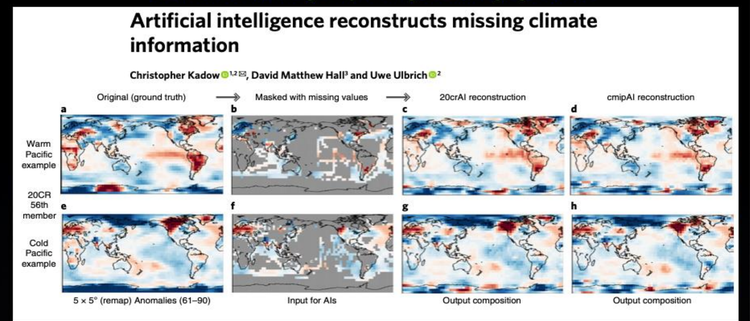


FILLING-IN MISSING CLIMATE OBSERVATIONS

DL transfer learning + inpainting beats Kriging and PCA

Artificial intelligence reconstructs missing climate information

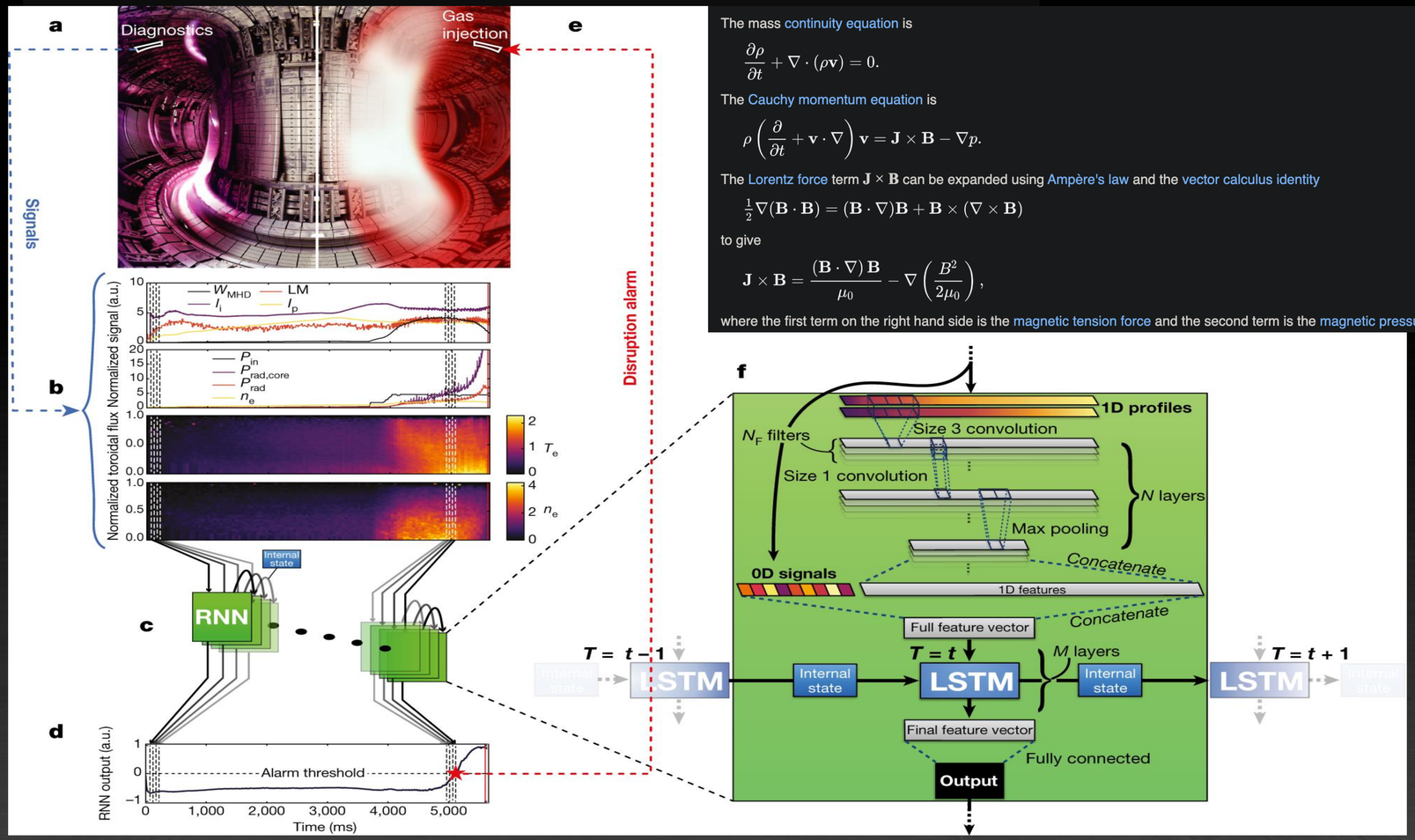
Christopher Kadow[1,2✉], David Matthew Hall[3] and Uwe Ulbrich[2]



Predicting disruptive instabilities in controlled fusion plasmas through deep learning

Julian Kates-Harbeck,[1,2] Alexey Svyatkovskiy,[3] William Tang[2,3]

[1]Department of Physics, Harvard University,
17 Oxford St, Cambridge, MA 02138, USA
[2]Princeton Plasma Physics Laboratory, Princeton, NJ, USA
[3]Princeton University, Princeton, NJ, USA



The mass continuity equation is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0.$$

The Cauchy momentum equation is

$$\rho \left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) \mathbf{v} = \mathbf{J} \times \mathbf{B} - \nabla p.$$

The Lorentz force term $\mathbf{J} \times \mathbf{B}$ can be expanded using Ampère's law and the vector calculus identity

$$\tfrac{1}{2}\nabla(\mathbf{B} \cdot \mathbf{B}) = (\mathbf{B} \cdot \nabla)\mathbf{B} + \mathbf{B} \times (\nabla \times \mathbf{B})$$

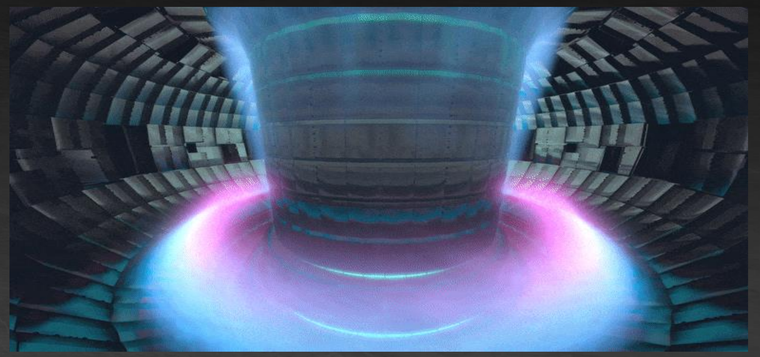
to give

$$\mathbf{J} \times \mathbf{B} = \frac{(\mathbf{B} \cdot \nabla)\,\mathbf{B}}{\mu_0} - \nabla \left(\frac{B^2}{2\mu_0} \right),$$

where the first term on the right hand side is the magnetic tension force and the second term is the magnetic pressure force.







Article | Open Access | Published: 16 February 2022

Magnetic control of tokamak plasmas through deep reinforcement learning

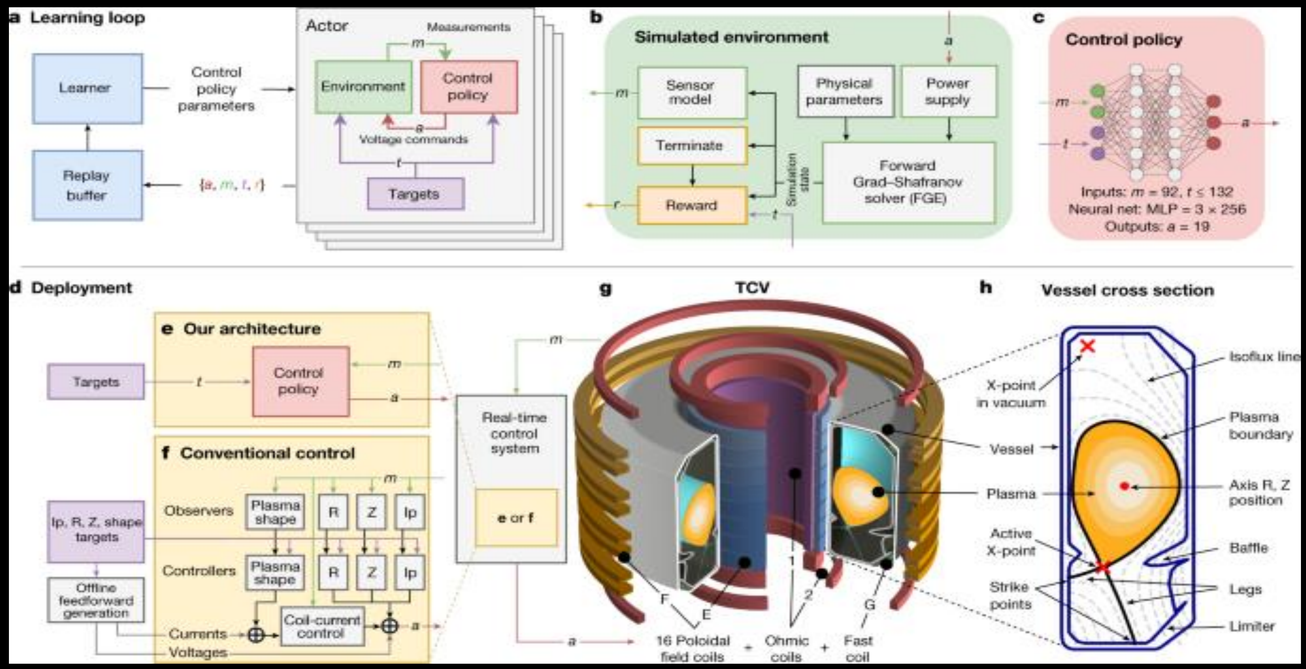
Jonas Degrave, Federico Felici ✉, ... Martin Riedmiller + Show authors

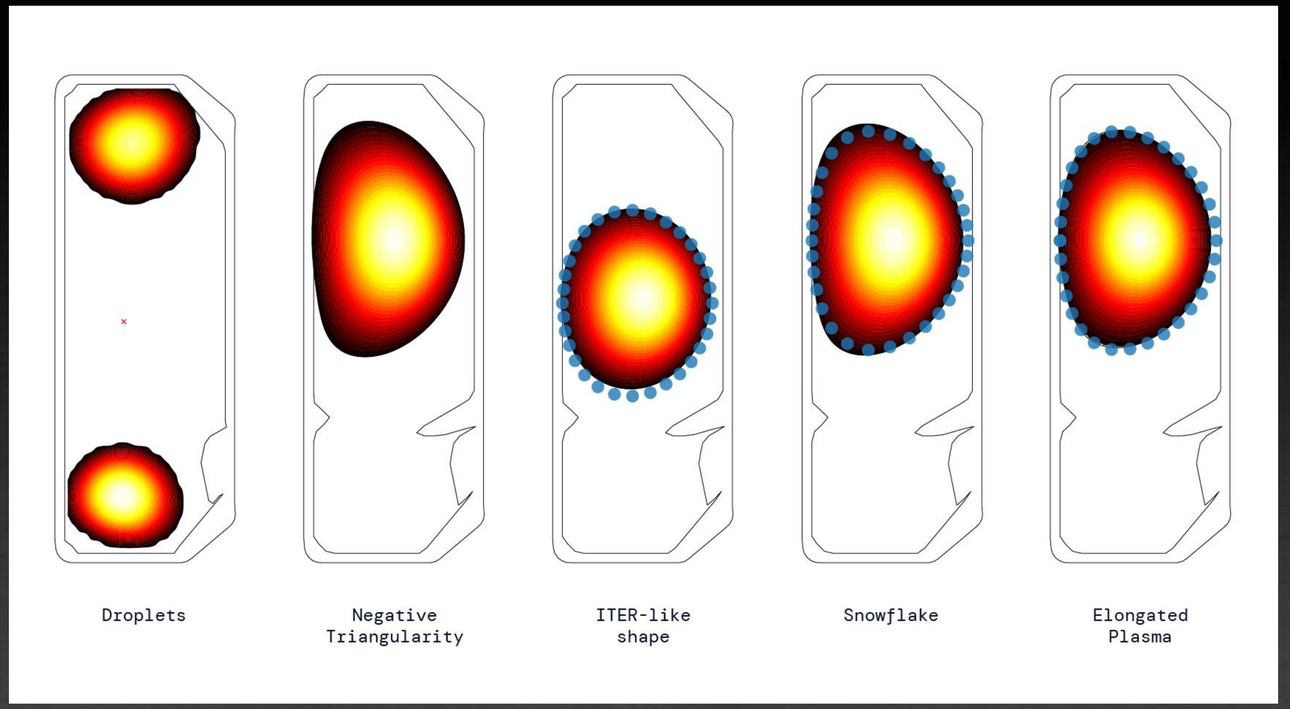
Nature **602**, 414–419 (2022) | Cite this article

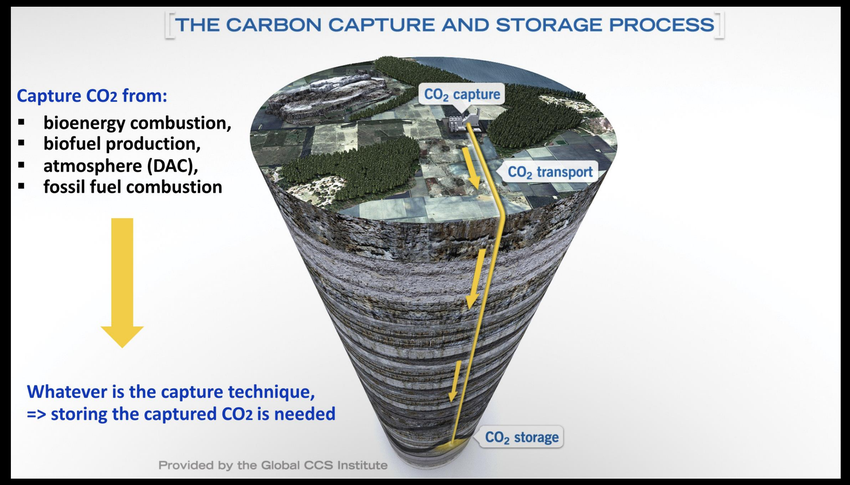
64k Accesses | **2042** Altmetric | Metrics

Abstract

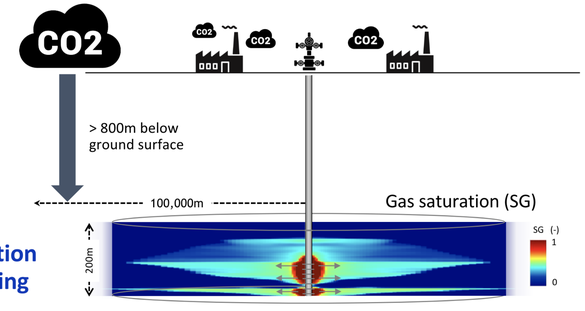
Nuclear fusion using magnetic confinement, in particular in the tokamak configuration, is a promising path towards sustainable energy. A core challenge is to shape and maintain a high-temperature plasma within the tokamak vessel. This requires high-dimensional, high-frequency, closed-loop control using magnetic actuator coils, further complicated by the diverse requirements across a wide range of plasma configurations. In this work, we introduce a previously undescribed architecture for tokamak magnetic controller design that autonomously learns to command the full set of control coils. This architecture meets control objectives specified at a high level, at the same time satisfying physical and operational constraints. This approach has unprecedented flexibility and generality in problem specification and yields a notable reduction in design effort to produce new plasma configurations. We successfully produce and control a diverse set of plasma configurations on the Tokamak à Configuration Variable[1,2], including elongated, conventional shapes, as well as advanced configurations, such as negative triangularity and 'snowflake' configurations. Our approach achieves accurate tracking of the location, current and shape for these configurations. We also demonstrate sustained 'droplets' on TCV, in which two separate plasmas are maintained simultaneously within the vessel. This represents a notable advance for tokamak feedback control, showing the potential of reinforcement learning to accelerate research in the fusion domain, and is one of the most challenging real-world systems to which reinforcement learning has been applied.





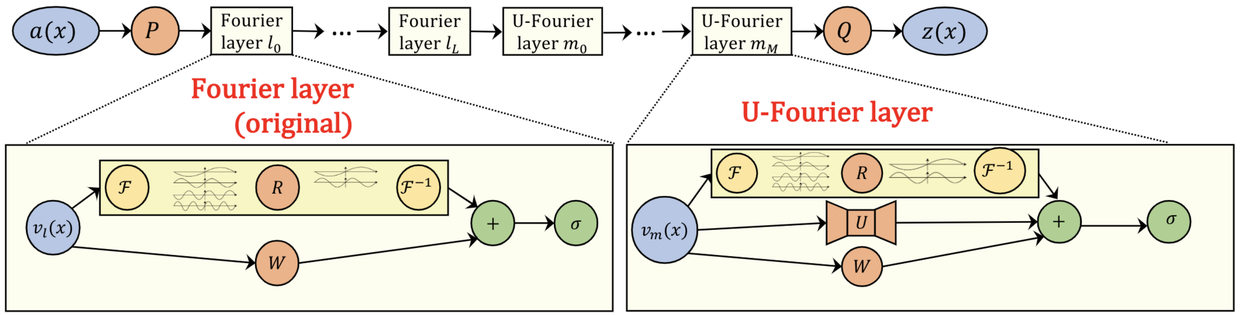


1. Capture CO_2
2. Transport CO_2
3. Inject CO_2 into subsurface
4. Store CO_2 in a saline aquifer reservoir
5. Model CO_2 plume migration and pressure buildup using AI methods



New

U-FNO ARCHITECTURE: ENHANCED FOURIER NEURAL OPERATOR



The U-FNO architecture appends a U-Net path in each U-Fourier layer

Characteristics of the U-Fourier Layer:

- ✓ 2-step mini U-Net to enhance higher frequency information
- ✓ The number of Fourier and U-Fourier layers, L and M, are hyperparameters, optimized for the specific problem

COMPUTATIONAL EFFICIENCY

CNN vs **FNO** vs **Conv-FNO** vs **U-FNO**

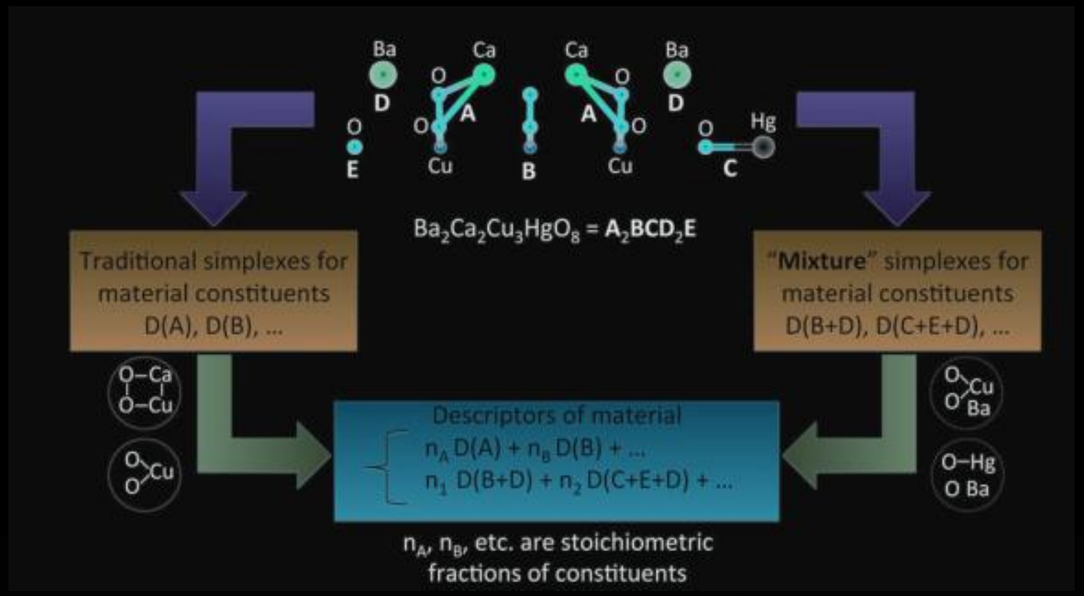
The speedup is compared with average numerical simulation runtime of 10 mins

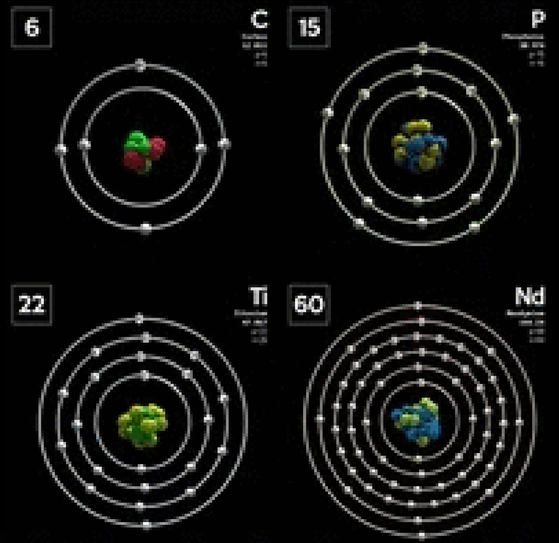
	# Parameter (-)	Training (s/epoch)	Testing		
			Gas saturation (s)	Pressure buildup (s)	Speed-up vs. numerical simulation (times)
CNN	33,316,481	562	0.050	0.050	1×10^4
FNO	31,117,541	711	0.005	0.005	1×10^5
Conv-FNO	31,222,625	1,135	0.006	0.006	1×10^5
U-FNO	33,097,829	1,872	0.010	0.010	6×10^4

U-FNO prediction speed up is 6 x 10^4 vs. FDM; even faster than CNN

The training and testing times are both evaluated on a Nvidia A100-SXM GPU and compared to ECLIPSE simulations on an Intel® Xeon® Processor E5-2670 CPU







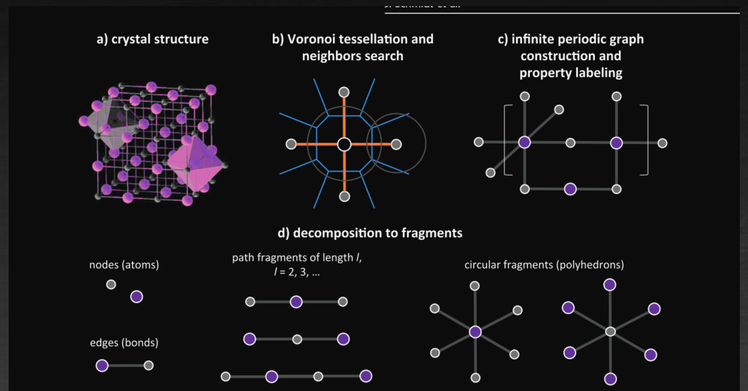
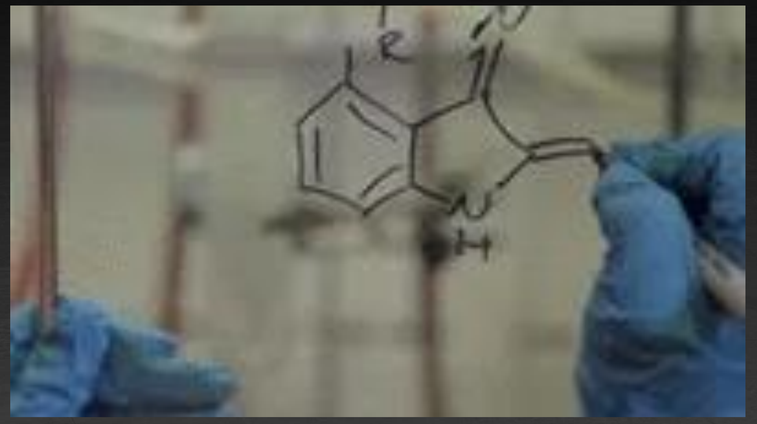


Fig. 5 Representation of the construction of property-labeled material fragments. The atomic neighbors of a crystal structure (**a**) are found via Voronoi tessellation (**b**). The full graph is constructed from the list of connections, labeled with a property (**c**) and decomposed into smaller subgraphs (**d**). (Reprinted with permission from ref. [41] licensed under the CC BY 4.0 [https://creativecommons.org/licenses/by/4.0/])



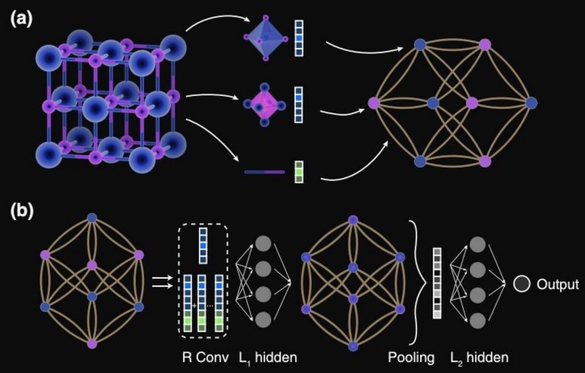
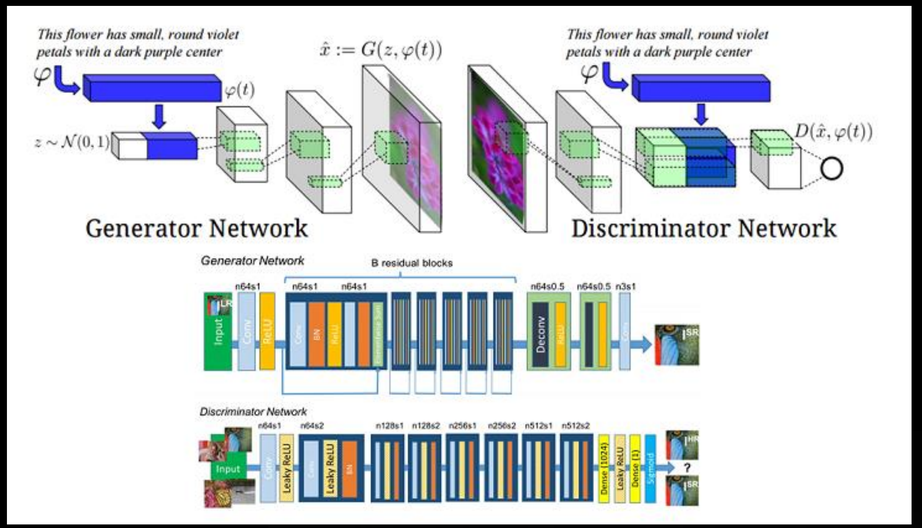


Fig. 6 Illustration of the crystal graph convolutional neural network. **a** Construction of the graph. **b** Structure of the convolutional neur network. (Reprinted with permission from ref. [40]. Copyright 2018 American Physical Society.)

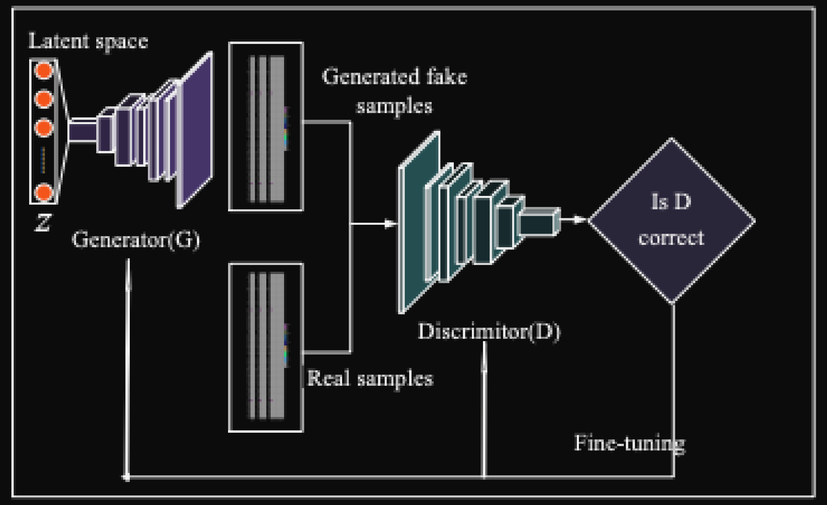
GENERATIVE ADVERSARIAL NETWORK





Generative adversarial networks (GAN) based efficient sampling of chemical space for inverse design of inorganic materials

Yabo Dan[1] Yong Zhao[2] Xiang Li[1] Shaobo Li[1,4] Ming Hu[3] and Jianjun Hu[1,2,*]



$$Loss_G = -\mathrm{E}_{x\sim P_g}\left[f_w(x)\right] \quad (1)$$

$$Loss_D = \mathrm{E}_{x\sim P_g}\left[f_w(x)\right] - \mathrm{E}_{x\sim P_r}\left[f_w(x)\right] \quad (2)$$

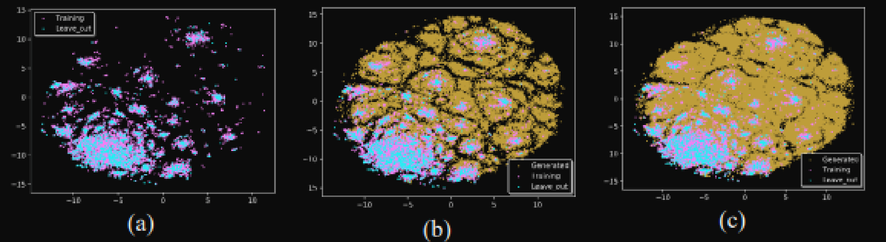


Fig. 3. Inorganic materials space composed of existing ICSD materials and hypothetical materials generated by GAN-ICSD. The two axes correspond to the two dimensions after t-sne based dimension reduction. The ICSD materials only occupies a tiny portion of the chemical space of inorganic materials. (a) Training samples (green dots) and leave-out validation samples (red dots) from ICSD; (b) 50,000 generated samples (blue dots) together with training and leave-out samples (c) 200,000 generated samples together with training and leave-out samples.

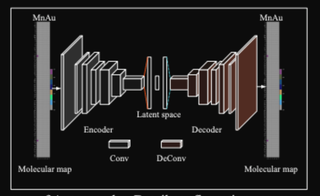


Fig. 2 Architecture of Autoencoder. Detail configuration parameters are shown in supplementary Table S3 and supplementary Fig S2.

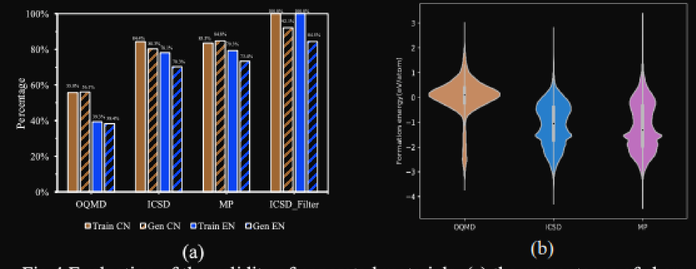


Fig.4 Evaluation of the validity of generated materials. (a) the percentages of charge-neutral (CN) and electronegativity-balanced (EN) samples of the generated samples are very close to those of the training sets for all four datasets. Train/gen CN: percentage of training/generated samples that satisfy charge neutrality; Train/gen EN: percentage of samples that satisfy balanced electronegativity. (b) Formation energy distribution of the Li-containing compounds generated by three GANs. Both GAN-ICSD and GAN-MP can generate a large percentage of hypothetical materials with low (< 0) formation energy.

NVIDIA AI TECHNOLOGY CENTER (NVAITC)

Catalyse AI transformation through research-centric integrated engagements







