AI for Science

Rick Stevens
Argonne National Laboratory
The University of Chicago

Crescat scientia; vita excolatur
Can a Google search prove that we've been visited by time travelers from the future?

A pair of physicists think there may be a way to test whether we've been visited by time travelers – by doing a simple Google search.

BRYAN NELSON  
January 9, 2014, 12:21 a.m.

Photo: William Warby/Flicker
Five Years Out

- Aurora and Frontier are running well and support a broad suite of large-scale applications in simulation, data analysis and machine learning.
- Progress is made on integrating machine learning into applications involving search and design such as Materials, Drugs, Biological Engineering, and Manufacturing Processes.
- USA National initiatives in AI and QC are running.
- Progress on the broad application of AI in Science and Engineering especially for Energy and Biomedical applications.
- We have well functioning testbeds for Quantum Computing and Neuromorphic Computing used by many users.
- Software stacks and programming environments that support integration of QC with Classical computing systems are being developed and are expected to be central for procurements of the first large-scale hybrid systems (vendors are behind in this integrative thinking).
Ten Years Out -- fuzzier

- Argonne and Oak Ridge have deployed a 100 Exaflops scale hybrid computers that has a mixture of compute devices, AI for science is now the driving application, and simulations are nearly all managed by AI
- ZettaOPS systems are being designed though its not clear how to account for the quantum, AI and neuromorphic compute elements. LinpackQ is hotly debated.
- AI is used to manage systems, augment programming environments, drive performance optimization and passively analyze all datasets, integrating ML and symbolic computing is going slowly
- Quantum computers have reached a million qubits and are now in wide use for selected simulation problems, QC programming is becoming mainstream in CS and languages such as Python 4, Julia Q, and FortranQ all support QC extensions. Quantum accelerated libraries are in wide use, though mostly by the AIs
- Partnering has become the normal way for US agencies and countries to gain access and expertise in computing AI, QC and represents a significant and by now is a regular pattern of collaboration
Fifteen Years Out -- much fuzzier

• ZettaOPS systems are being deployed, the non-von computing elements dominate the compute power, cost and development

• Storage systems have been replaced with in situ AI that provide multi-level abstractions for accessing the information content from raw bits to learned functions that capture distributions

• Nearly everything that collects data has some intelligence layered on it and its routine to access systems via voice and gesture interfaces

• Quantum computers that use distributed entanglement are now in wide use, with systems architectures able to scale to $10^9$ qubits

• A major use of simulation is to augment training datasets for filling in gaps of multi-task AIs in science and engineering. It's now routine to widely share models rather than data.

• Foundational work on the quantum, biological and molecular basis for future computing is accelerating and investments in quantum materials and synthetic biology are pushing frontiers of computing architecture
An AI plan for DOE
or
a roadmap for post-exascale computing?
**AI for Science Townhalls**

Organized by Argonne, Oak Ridge and Berkeley with participation from all the laboratories.

- Four “Townhalls” aimed at getting input from the DOE community on opportunities and requirements for the next 5-10 years in computing with a focus on convergence between HPC and AI
- July (Argonne), August (Oak Ridge), September (Berkeley), October (Washington)
- Modeled after the 2007 Townhalls that launched the Exascale Computing Initiative
- Each meeting covers roughly the same ground, geographically distributed to enable local participation
- Applications in science, energy and technology
- Software, math and methods, hardware, data management, computing facilities, infrastructure, integration with experimental facilities, etc.
- Expect ~200-350 people per meeting
- Output will be a report to guide strategic planning at Labs and DOE
AI for Science Town Hall Series Kicks off at Argonne

By Tiffany Trader
What we are on about at the Townhalls

- AI is transforming our “regular life” world
- AI has tremendous potential to accelerate scientific discovery
- How do we go about organizing an AI for Science initiative
- Capture ideas, problems, requirements and challenges for an AI for Science initiative
- What problems could be attacked?
- What data, simulations, and experiments do we need?
- What kind of methods, software and math do we need?
- What kind of computer architectures and infrastructure do we need?
AI complements our Exascale plans

- The emerging platforms at the LCF and NERSC will be excellent platforms for machine learning, in particular deep learning training
- The coupling of AI and HPC is a huge opportunity for DOE
- Many uses of AI couple to experiments in ways that traditional modeling and simulation do not
- The DOE experimental community could become major users of the DOE HPC facilities
- AI has the potential to accelerate science at all scales
- Future systems directions will be impacted by AI use cases
What do we mean by AI
Human Vision System
Deep Learning impact on Image classification research

**Figure 4.** Historical error rate of the best performing image classification algorithms in the annual ImageNet competition. Established models of computer vision stagnated at 25–30%. The introduction of deep learning in 2012 led to a significant improvement to ~15%, and human-level accuracy (~5%) for image classification was achieved by 2015.
### What Machine Learning Can Do

A simple way to think about supervised learning.

<table>
<thead>
<tr>
<th>INPUT A</th>
<th>RESPONSE B</th>
<th>APPLICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture</td>
<td>Are there human faces? (0 or 1)</td>
<td>Photo tagging</td>
</tr>
<tr>
<td>Loan application</td>
<td>Will they repay the loan? (0 or 1)</td>
<td>Loan approvals</td>
</tr>
<tr>
<td>Ad plus user information</td>
<td>Will user click on ad? (0 or 1)</td>
<td>Targeted online ads</td>
</tr>
<tr>
<td>Audio clip</td>
<td>Transcript of audio clip</td>
<td>Speech recognition</td>
</tr>
<tr>
<td>English sentence</td>
<td>French sentence</td>
<td>Language translation</td>
</tr>
<tr>
<td>Sensors from hard disk, plane engine, etc.</td>
<td>Is it about to fail?</td>
<td>Preventive maintenance</td>
</tr>
<tr>
<td>Car camera and other sensors</td>
<td>Position of other cars</td>
<td>Self-driving cars</td>
</tr>
</tbody>
</table>

SOURCE: ANDREW NG
Why deep learning

Performance

Amount of data

Deep learning
Older learning algorithms
We see this effect of model capacity in our work on Cancer drug response prediction.
Thing Translator

A.I. Experiments:

Thing Translator
Machine Learning Arxiv Papers per Year

- ML Arxiv Papers
- Moore's Law growth rate (2x/2 years)

~100 new ML papers every day!
What is possible?
Things we can do in Science with AI now

Learn predictive models from data without relying upon theory or deep mechanistic understanding

   Example: predicting materials and chemistry properties

Learn approximate solutions to inverse problems where we have data and models are not available or are inefficient

   Example: phase retrieval in coherent x-ray imaging

Generate large collections of synthetic data that models real data

   Example: synthetic sky in cosmology
Deep Learning in Chemistry

- Drug Design
- Catalysis Design
- Materials Design
- Designing Molecules
- Synthesizing Molecules
- QSPR
- Conformer Exploration
- Accelerated QM
- Retro-synthesis
- Reaction Prediction
- Reaction Optimization
Figure 1. The Big Picture of Deep Learning. The learner shown in this image is a deep feedforward network; however, this same procedure applies to a plethora of learners. The ΔP term indicates the change to the parameters in each network layer after the input layer. The data in this image is fictitious and thus labeled simply as property.
Modeling Cancer Drug Response

\[ R = f(T, D_1, D_2) \]

- Drug(s)
  - descriptors
  - fingerprints
  - structures
  - SMILES
  - dose

- Gene expression levels
- SNPs
- Protein abundance
- microRNA
- Methylation

IC50
AUC
GI50
% growth
Z-score

Response

Tumor

Drug Concentration in Log scale

Response

U.S. DEPARTMENT OF ENERGY
NIH NATIONAL CANCER INSTITUTE
“Uno” Model Predictions with Dropout UQ (trained on ALMANAC)

All Samples colored by Sample ID  
Samples found in Cluster 1 or Cluster 2 colored by Sample ID

Pilot 1

RTS subset of drug pairs

x NCIPDM.237351~077-R~AL-IR0
x NCIPDM.994434~217-R~AK3YH7
x NCIPDM.CN0446~F447~M12M52
Things We Want To Do With AI In The Future

• Develop methods that can learn from both encoded symbolic theory (e.g. QM/GR) and large-scale data so we can leverage the vast theoretical knowledge we have accumulated over hundreds of years.

• Automate and accelerate discovery from planning, to conjecture, to experiment, to confirmation and analysis ⇒ end-to-end automated science.

• Create an ability to use AI for generating new theories that address the problematical areas of existing theories.
In Ten Years…

- **Learned Models Begin to Replace Data**
  - queryable, portable, pluggable, chainable, secure

- **Experimental Discovery Processes Dramatically Refactored**
  - models replace experiments, experiments improve models

- **Many Questions Pursued Semi-Automatically at Scale**
  - searching for materials, molecules and pathways, new physics

- **Simulation and AI Approaches Merge**
  - deep integration of ML, numerical simulation and UQ

- **Theory Becomes Data for Next Generation AI**
  - AI begins to contribute to advancing theory

- **AI Becomes Common Part of Scientific Laboratory Activities**
  - Infuses scientific, engineering and operations
A Sampling of Science Opportunities
Materials and Chemistry

- Design of materials and molecules
- Al-guided synthesis
  - automated design of chemical pathways
  - mapping metastable phases
  - extracting mechanisms
- Predictive interfacial transport of ions and charge
- Al-accelerated ab Initio molecular dynamics
- Quantification of energy drivers for separations
- Describing multiscale charge, spin, lattice correlations
- Exploring energy landscapes in ultrafast, nonequilibrium, and driven systems and processes
- Inverse design, bandstructure engineering
ML models achieve impressive results for many materials problems

Table 1  Materials informatics model results from the literature. The Pearson correlation coefficient $R$ between predicted and actual property values is a common means of quantifying model performance. RMSE is root mean square error; MAE is mean absolute error; $R^2$ is the square of the Pearson correlation coefficient.

<table>
<thead>
<tr>
<th>Material class</th>
<th>Property</th>
<th>ML technique</th>
<th>CV type</th>
<th>Model performance metric</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steel</td>
<td>Fatigue strength</td>
<td>Multivariate polynomial regression</td>
<td>Leave-one-out CV</td>
<td>$R^2 = 0.9801$</td>
<td>3</td>
</tr>
<tr>
<td>Organic small molecules</td>
<td>Norm of dipole moment</td>
<td>Graph convolutions</td>
<td>Overall 90% train/10% test, with reported test error averaged across 10 different models built on subsets of training data</td>
<td>MAE = 0.101 Debye (chemical accuracy target: 0.10 Debye)</td>
<td>4</td>
</tr>
<tr>
<td>Polymers</td>
<td>Electronic dielectric constant</td>
<td>Kernel ridge regression</td>
<td>81% train/19% test</td>
<td>$R^2 = 0.96$</td>
<td>16</td>
</tr>
<tr>
<td>Inorganic compounds</td>
<td>Formation energy</td>
<td>Rotation forest</td>
<td>32% train/68% test</td>
<td>$R^2 = 0.93$</td>
<td>5</td>
</tr>
<tr>
<td>Inorganic compounds</td>
<td>Vibrational free energy</td>
<td>Random forest or support vector machine</td>
<td>10 averaged $k$-fold CV runs, for $k$ in [ref. 5 and 14]</td>
<td>$R = 0.95$</td>
<td>6</td>
</tr>
<tr>
<td>Inorganic compounds</td>
<td>Band gap</td>
<td>Support vector machine</td>
<td>100 averaged 75% train/25% test runs</td>
<td>$G_0W_0$ RMSE = 0.18 eV (DFT RMSE $\sim$ 2 eV wrt exp.)</td>
<td>7</td>
</tr>
</tbody>
</table>

Meredig et al., DOI:10.1039/c8me00012c
Climate and Biology

- Accelerated Climate Models (PDE/ML hybrids)
- Improved integration of remote sensing and ground truthing into Climate Models (cloud/precipitation, land cover/biogeochem, sea ice/calibration, etc.)
- Improvement in ARM data pipelines, automated model extraction from data, smart data fusion

- Vast applications in genomics and metagenomics (G ⟷ P)
- Automation of bioinformatics methods (improved productivity)
- Automating hypothesis formation in biology (causal analysis)
- Forward design of novel pathways, proteins, regulons, operons, organisms, etc. for secure biodesign
- Anomaly detection (discovery in sequencing, biosecurity, etc.)
A Network that Learns Drug-Drug Interactions

Fig. 1  KMR for drug representation learning and drug-drug interaction prediction. Red, green, blue and yellow matrices denote pharmacological feature representations, drug textual description feature representations, drug class feature representations, and final knowledge-oriented drug representations, separately.
High Energy Physics

Energy/Intensity Frontier:
- Search for Beyond the Standard Model (BSM) physics through AI-driven anomaly detection
- AI-reduced uncertainties to enable precision electroweak measurements for BSM clues
- Generative Adversarial Networks (GANs) for large-scale Large Hadron Collider detector simulation

Cosmic Frontier – AI in end-to-end application:
- Precision Cosmic Microwave Background emulation – AI simulation speed-up of a factor of 1000
- Search for strong lensing of galactic sources for precision cosmology measurements using AI classification, regression, and GANs for image generations
- AI-based Photometric Redshift Estimation
- Combination of AI methods to enable searches for hidden space variables

AI applications in an “end-to-end” Cosmic Frontier application: 1) GANs for image emulation, 2) GP and DL-based emulators for summary statistics, 3) CNN-based image classification, 4) AI-based photometric redshift estimation, 5) Likelihood-free methods for inference [Work performed under the Argonne-led SciDAC-4 project: “Inference and Machine Learning at Extreme Scales”]
3D convolutional GAN

- Similar discriminator and generator models
- 3D convolutions (keep X,Y symmetry)
- Tested several tips&tricks from literature
- Some helpful (no batch normalisation in the last step, LeakyRelu, no hidden dense layers, no pooling layers)
- RMSProp optimiser for both networks
- Batch training

(*) https://github.com/soumith/ganhacks
### Computing performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Machine</th>
<th>Time/Shower (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full Simulation (geant4)</strong></td>
<td>Intel Xeon Platinum 8180</td>
<td>17000</td>
</tr>
<tr>
<td>3d GAN (batch size 128)</td>
<td>Intel Xeon Platinum 8180</td>
<td>7</td>
</tr>
<tr>
<td>3d GAN (batchsize 128)</td>
<td>GeForce GTX 1080</td>
<td>0.04</td>
</tr>
</tbody>
</table>

- Inference: speedup factor > 2500

---

#### Training (speedup vs. performance)

- 45 min/epoch on Tesla P100
- Introduce data parallel training based on MPI
- Test several libraries
- More in the MS04 mini-symposium
AI and HPC
Connecting HPC and AI

In addition to partnerships in AI applications, there are considerable opportunities in foundational methods development, software and software infrastructure for AI workflows and advanced hardware architectures for AI, below we highlight some ideas in the HPC + AI space

- Steering of simulations
- Embedding simulation into ML methods
- Customized computational kernels
- Tuning applications parameters
- Generative models to compare with simulation
- Student (AI) Teacher (Sim) models $\Rightarrow$ learned functions
- Guided search through parameter spaces
- Hybrid architectures HPC + Neuromorphic
- Many, many more
Robust Learned Function Accelerators
FIG. 2. In this work, we use the machinery of deep learning to learn the mapping between potential and energy, bypassing the need to numerically solve the Schrödinger equation and the need for computing wave functions. The architecture we used (shown here) consisted primarily of convolutional layers capable of extracting relevant features of the input potentials. Two fully connected layers at the end serve as a decision layer, mapping the automatically extracted features to the desired output quantity. No manual feature selection is necessary; this is a featureless-learning approach.
Deep learning and the Schrödinger equation

Kyle Mills, Michael Spanner, and Isaac Tamblyn
Phys. Rev. A 96, 042113 – Published 18 October 2017

FIG. 4. Histograms of the true vs predicted energies for each example in the test set indicate the performance of the various models: (a) simple harmonic oscillator, (b) infinite well, (c) DIG potential, (d) random potential, and (e) DIG potential on random model. The insets show the distribution of error away from the diagonal line representing perfect predictions. A 1-mHa² square bin is used for the main histograms and a 1-mHa bin size for the inset histogram. During training, the neural network was not exposed to the examples on which these plots are based. The higher error at high energies in (d) is due to fewer training examples being present in the data set at these energies. The histogram shown in (d) is for the further-trained model, described in the text.

FIG. 11. Examples of the four classes of potentials.
Figure 5.2 Comparison between speed and accuracy of neural network versus DFT in energy prediction for organic molecules.
AI at Argonne: Broad Span of Scientific Targets

- Reduced order modeling of laser sintering
- Nowcasting with convolutional LSTMs
- Prediction of radiation stopping power
- Strong and weak lensing in sky survey data
- Prediction of antimicrobial resistance phenotypes
- Identification and tracking of storms
- Efficient climate model emulators
- Defect-level prediction in semiconductors
- Structure-property-process triangle in additive manufact.
- Parameter extraction in atom probe tomography
- Learning for dynamic sampling in spectroscopy
- Vehicle energy consumption prediction
- Flying object detector for edge deployment
- Discovery of new energy storage materials
- Cosmic Microwave Background emulation
- Photometric red shift estimation
- New materials for efficient solar cells
- Enhancement of noisy tomographic images
Example from Traumatic Brain Injury
Anatomical Segmentation

- Caudate
- Pallidum
- Putamen
- Amygdala
- Hippocampus
- White Matter
- Cortex
- Lateral Ventricle
- Thalamus

Connectomics

- Original Runtimes
- March timings
- Current timings
- Patient Scan Preprocessing
- Connectome Analysis

Diagnosis/Prognosis

- CT Scan
- DL Model
- 85% accurate

CT ↔ MRI “Super-resolution”

CT

MRI

Patient working after 6 months?
Training with diverse data modalities and phenotypes
Enhance CT imaging and exploit labels from other modalities

Generative Adversarial Networks
GAN Model trained on TBI patient data
Diverse brain disease MRI data for identifying abnormal CT

CNN Model trained on normal/abnormal MRI slices

Normal MRI

Tumor
- Meningioma
- Glioma

Stroke Lesion

Knowledge transfer for CT

Normal CT

TBI Lesion/Midline Shift
Building the AI Environment for Science
DLHub: Organizing and Serving Models

https://www.dlhub.org

- Collect, publish, categorize models
- Serve models via API with access controls to simplify sharing, consumption, and access
- Leverage ALCF resources and prepare for Exascale ML
- Deploy and scale automatically
- Provide citable DOI for reproducible science

Argonne Advanced Computing LDRD
CANDLE: Exascale Deep Learning Tools

Deep Learning Needs Exascale
- Automated model discovery
- Hyper parameter optimization
- Uncertainty quantification
- Flexible ensembles
- Cross-Study model transfer
- Data augmentation
- Synthetic data generation
- Reinforcement learning

https://github.com/ECP-CANDLE
CANDLE Project

- **CANDLE Python Library** – make it easy to run on DOE Big Machines, scale for HPO, UQ, Ensembles, Data Management, Logging, Analysis
- **CANDLE Benchmarks** – exemplar codes/models and data representing the three primary challenge problems
- **Runtime Software** – Supervisor, Reporters, Data Management, Run Data Base
- **Tutorials** – Well documented examples for engaging the community
- **Contributed Codes** – Examples outside of Cancer, including Climate Research, Materials Science, Imaging, Brain Injury
- **Frameworks** – Leverage of TensorFlow, Keras, Horovod, PyTorch, etc.
- **LL Libraries** – CuDNN, MKL, etc. (tuned to DOE machines)

**ECP-CANDLE GitHub Organization:**

[https://github.com/ECP-CANDLE](https://github.com/ECP-CANDLE)
Methods Innovation, one page agenda 😊

- Data efficient learning. “Low data”. One shot, few shot learning.
- Improved neural architecture search. Lottery tickets and sparsity.
- Online learning and incremental training. Active learning.
- Representation learning in novel “scientific” spaces.
- UQ and confidence estimates. Interpretability.
- Beyond NLP and CV towards concepts directly needed by science.
- Generative methods in scientific and engineering domains.
- Inverse methods and systems that input data and output rules.
Deep learning is great, but it requires too much data to be the long term solution
It seems that the advent of models that beat the power-law exponent — that get **more data efficient as they learn** — might be an important empirical milestone on that path.

What Can we Learn from Trained Models?
Fig. 2  Schematic of CNN model structure. The model takes the Raman spectrum and the composition as input to predict $P$. The differently colored layers correspond to red: dense layers acting on composition, green: convolutional 1D layers acting on spectra, yellow: flattening and concatenation layers, blue: dense layers acting on both the composition and spectral data. Each of the 10 layers of the CNN model are labelled a to j.
Diving into the gradients of the network

(npj Computational Materials (2019)5:34; https://doi.org/10.1038/s41524-019-0172-5)
<table>
<thead>
<tr>
<th>#</th>
<th>Data observations from analysis of CNN models</th>
<th>Human-derived materials science explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adding &lt;A&gt; in the high-P samples increase P by an average of &lt;B&gt; mW cm²/ at. %</td>
<td>Synergy of both Mo-Tb and Mo-Gd co-alloying lead to beneficial structure modulations</td>
</tr>
<tr>
<td>2</td>
<td>Mo,Tb,Gd</td>
<td>0.25,0.10,0.10</td>
</tr>
<tr>
<td>3</td>
<td>(Dy, Gd, Tb)</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>(W, Mo)</td>
<td>0.91</td>
</tr>
<tr>
<td>5</td>
<td>For increasing conc. of &lt;A&gt;, &lt;B&gt; Raman intensity in &lt;C&gt; cm⁻¹ has similar influence on P.</td>
<td>All 3 REs have the same role of decreasing mono. dist.</td>
</tr>
<tr>
<td>6</td>
<td>Dy,Gd,Tb</td>
<td>Decreasing 331.4–332.7</td>
</tr>
<tr>
<td>7</td>
<td>Dy,Gd,Tb</td>
<td>Decreasing 359.1–361.7</td>
</tr>
<tr>
<td>8</td>
<td>W,Mo</td>
<td>Increasing 339.3–357.7</td>
</tr>
<tr>
<td>9</td>
<td>Dy,Gd,Tb</td>
<td>Decreasing 698.9–728.9</td>
</tr>
<tr>
<td>10</td>
<td>W,Mo</td>
<td>Increasing 791.9–796.9</td>
</tr>
<tr>
<td>11</td>
<td>W,Mo</td>
<td>Increasing 834.9–836.9</td>
</tr>
<tr>
<td>12</td>
<td>W,Mo</td>
<td>Decreasing 127.3–143.0</td>
</tr>
<tr>
<td>13</td>
<td>W,Mo</td>
<td>Decreasing 212.6–224.2</td>
</tr>
<tr>
<td>14</td>
<td>W,Mo</td>
<td>Increasing 880.0–884.8</td>
</tr>
<tr>
<td>15</td>
<td>W,Mo</td>
<td>Increasing 302.2–303.5</td>
</tr>
<tr>
<td>16</td>
<td>W,Mo</td>
<td>Increasing 667.6–668.9</td>
</tr>
</tbody>
</table>

For each type of observation, a human-interpretable sentence is provided with fill-in values noted for each enumerated observation. The right-most column is the human-generated explanation of the underlying materials science that gave rise to each data relationship, as discussed in our previous work and using literature identification of Raman modes.
Uncertainty Quantification... can models tell us how confident they are in their predictions?
Accuracy vs. Confidence for Test Set

200 Inference Trials per Sample in Test Set

High density of predictions encountered around zero error.

Different methods exhibit slightly different statistics.
UQ Empirical Calibration

1. Predict a distribution of values (e.g. bootstrap or dropout)
2. Use this predictive distribution to compute relation between accuracy (error) and confidence (STD)

Known ground truth

Unknown ground truth

Empirical relation obtained from prediction statistics where ground truth is known

Use standard deviation of predictions and empirical calibration
Empirical Calibration: STD vs. Error

This plot highlights region of one-to-one relation between accuracy and confidence.

Since monotonically increasing predicted STD is not equal to accuracy, it is necessary to calibrate the predictions: i.e. estimate a mapping from confidence (STD) to accuracy (error).

This one-to-one mapping is used to establish the expected accuracy of new samples via their predicted STD.
We are starting out in a good place
Deep Learning Needs HPC

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute

ExaFLOPs for one Day
Aurora: HPC and AI

>> Exaops/s for AI

Architecture supports three types of computing
- Large-scale Simulation (PDEs, traditional HPC)
- Data Intensive Applications (scalable science pipelines)
- Deep Learning and Emerging Science AI (training and inferencing)
Specialized hardware is emerging that will be 10x – 100x the performance of general purpose CPU and GPU designs for AI.

US VCs investing >$4B in startups for AI acceleration.

Which platforms will be good for science?
<table>
<thead>
<tr>
<th>Tech Giants/Systems</th>
<th>IC Vender/Fabless</th>
<th>IP/Design Service</th>
<th>Startup in China</th>
<th>Startup Worldwide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Intel</td>
<td>ARM</td>
<td>Cambricon</td>
<td>cerebras</td>
</tr>
<tr>
<td>Microsoft</td>
<td>Samsung</td>
<td>Synopsys</td>
<td>Huawei</td>
<td>WAVE Computing</td>
</tr>
<tr>
<td>Amazon</td>
<td>NVIDIA</td>
<td>Imagination</td>
<td>Neurona</td>
<td>Graphcore</td>
</tr>
<tr>
<td>Facebook</td>
<td>Qualcomm</td>
<td>Cadence</td>
<td>BITMAIN</td>
<td>habana</td>
</tr>
<tr>
<td>Apple</td>
<td>AMD</td>
<td>CEVA</td>
<td>intellITusion</td>
<td>thinc</td>
</tr>
<tr>
<td>Huawei</td>
<td>NXP</td>
<td>VeriSilicon</td>
<td>ChipIntelli</td>
<td>SambaNova</td>
</tr>
<tr>
<td>Baidu</td>
<td>Xilinx</td>
<td>Arteris</td>
<td>ThinkForce</td>
<td>Habanero</td>
</tr>
<tr>
<td>Alibaba Group</td>
<td>Xilinx</td>
<td>Artis</td>
<td>Canacan</td>
<td>8th Sense</td>
</tr>
<tr>
<td>Toshiba</td>
<td>SiFive</td>
<td>Synopsys</td>
<td>Acleris</td>
<td>Esperanto</td>
</tr>
<tr>
<td>MediaTek</td>
<td>ARM</td>
<td>Synopsys</td>
<td>LogiCore</td>
<td>Preferred Networks</td>
</tr>
<tr>
<td>Fujitsu</td>
<td>MAXELL</td>
<td>nGraph Compiler</td>
<td>Rokid</td>
<td>brainchip</td>
</tr>
<tr>
<td>NOKIA</td>
<td>NXP</td>
<td>NVIDIA TensorRT</td>
<td>adipose</td>
<td>Peka</td>
</tr>
<tr>
<td>Toshiba</td>
<td>Rockchip</td>
<td>mlperf</td>
<td>Greenwaves</td>
<td>Greenwaves</td>
</tr>
<tr>
<td>Enterprise</td>
<td>NVIDIA</td>
<td>AI - Benchmark</td>
<td>Markable</td>
<td></td>
</tr>
<tr>
<td>Honeywell</td>
<td>mlperf</td>
<td>AI Matrix</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Compiler: Glow, LLVM, nGraph Compiler stack (Beta), mlperf

Benchmarks: AI - Benchmark, AI Matrix
Largest Chip Ever Built

- 46,225 mm² silicon
- 1.2 trillion transistors
- 400,000 AI optimized cores
- 18 Gigabytes of On-chip Memory
- 9 PByte/s memory bandwidth
- 100 Pbit/s fabric bandwidth
- TSMC 16nm process
AI Accelerator Testbed

Engaging the community to understand and improve specialized AI hardware for science

Dozens of proposed AI accelerators promise 10x - 1000x acceleration for AI workloads. AI testbed will:
1. Provide an open and unbiased environment for evaluation of AI accelerator technologies
2. Disseminate information about use cases, software, performance on test problems
3. Support collaborations with AI technology developers, academics, commercial AI, DOE labs

Staged evaluation enables identification of most promising systems for science

https://github.com/basicml/AI-Chip
AI Driven Experimental Science
The ATOM Platform
Active Learning Drug Discovery Framework

Machine Learning Pipeline
- Efficacy
- Safety
- PK
- Developability

Multi-Level Models
- Multi-level models
- Systems models

Multi-Parameter Optimization Loop
- Design Criteria

Generative Molecular Design
- Proposes new molecules with optimized properties

Simulation
- Active learning decides if/when a simulation or experiment is needed to improve or validate models

Experiment
- Human-relevant assays, complex in vitro models
- Chemistry Design & Synthesis

Working Compound Library
- Retrain property prediction models

Jim Brase (LLNL) and the ATOM Consortium
Layered workflow combining AI, HPC and HTS

- Filter Candidates ML
- ML Property Prediction Pipeline
- ML Generator of Candidates
- UQ Scoring and Optimization
- Update Models ML
- Simulation: Estimation of Properties
- Experiment: Estimation of Properties
- Active Learning Prioritization

Pure ML “constant time” (fast loop)  Mixed/Variable time (slow loop)
AI Driven Autonomous Laboratory Cluster

- Common AI/ML Methods
- Data and Computing Infrastructure
- Polymer Design
- Genetically Engineered Materials
- Energy Materials
- Protein Design
- Additive Manufacturing
- Organism Design
We are off to a good start..

But we gotta play it Cool.