“A Vision for Exascale: Simulation, Data and Learning”

Rick Stevens

Argonne National Laboratory

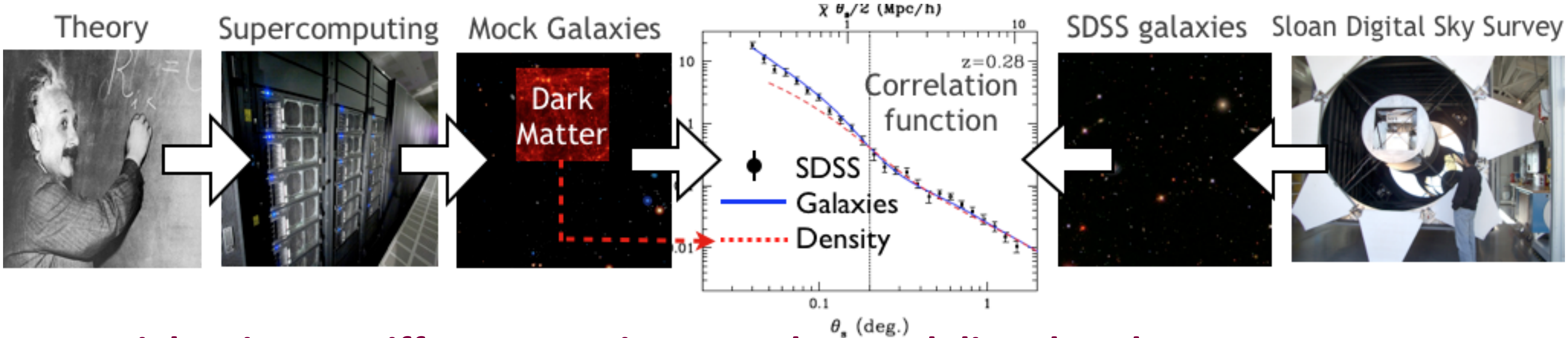
The University of Chicago

Crescat scientia; vita excolatur

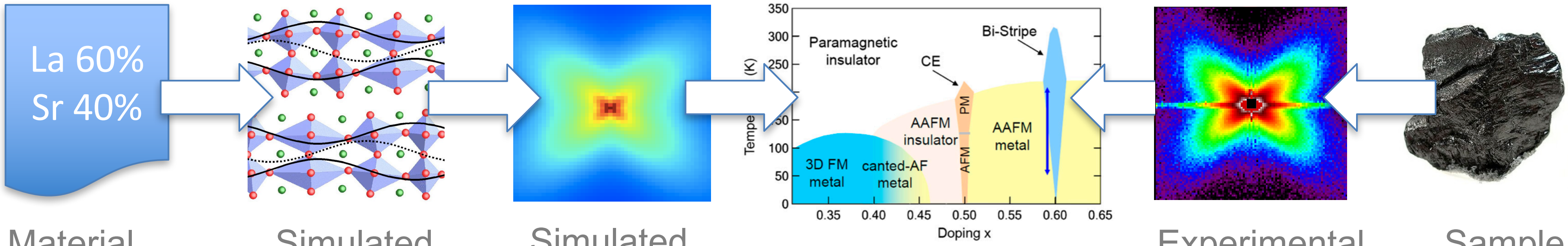
Data-Driven Science Examples

For many problems there is a deep coupling of observation (measurement) and computation (simulation)

Cosmology: The study of the universe as a dynamical system


Theory → Supercomputing → Mock Galaxies → Correlation function ← SDSS galaxies ← Sloan Digital Sky Survey

Materials science: Diffuse scattering to understand disordered structures



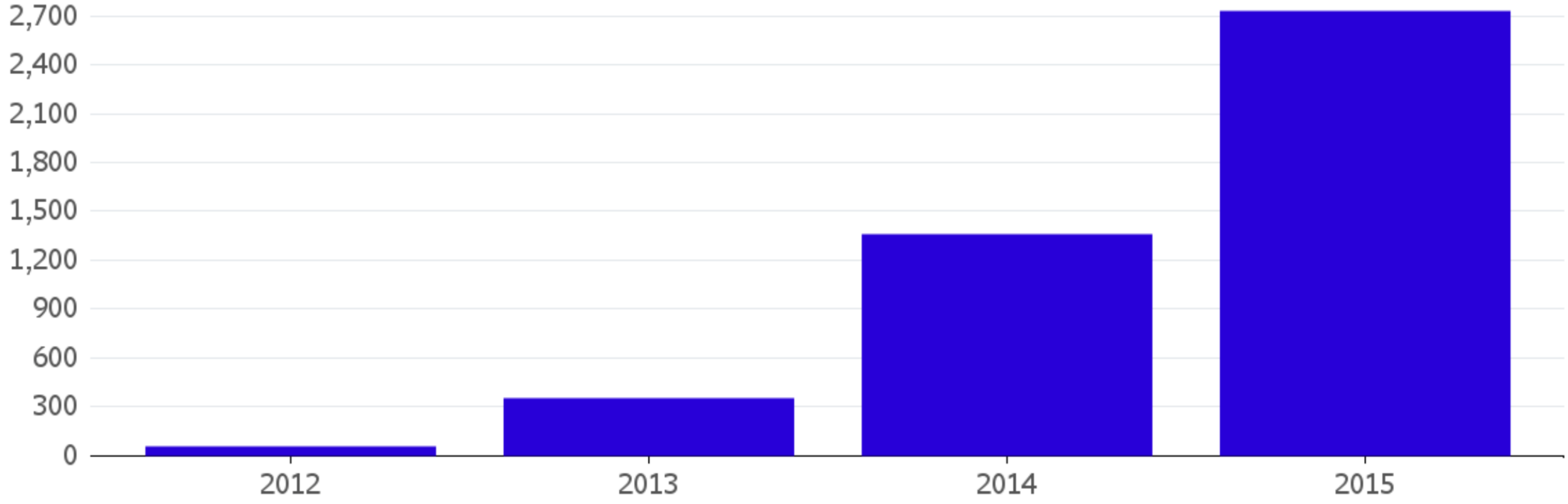
Material composition → Simulated structure → Simulated scattering → ← Experimental scattering ← Sample

Images from Salman Habib et al. (HEP, MCS, etc.) and Ray Osborne et al. (MSD, APS, etc.)

How Many Projects?

Artificial Intelligence Takes Off at Google

Number of software projects within Google that uses a key AI technology, called Deep Learning.



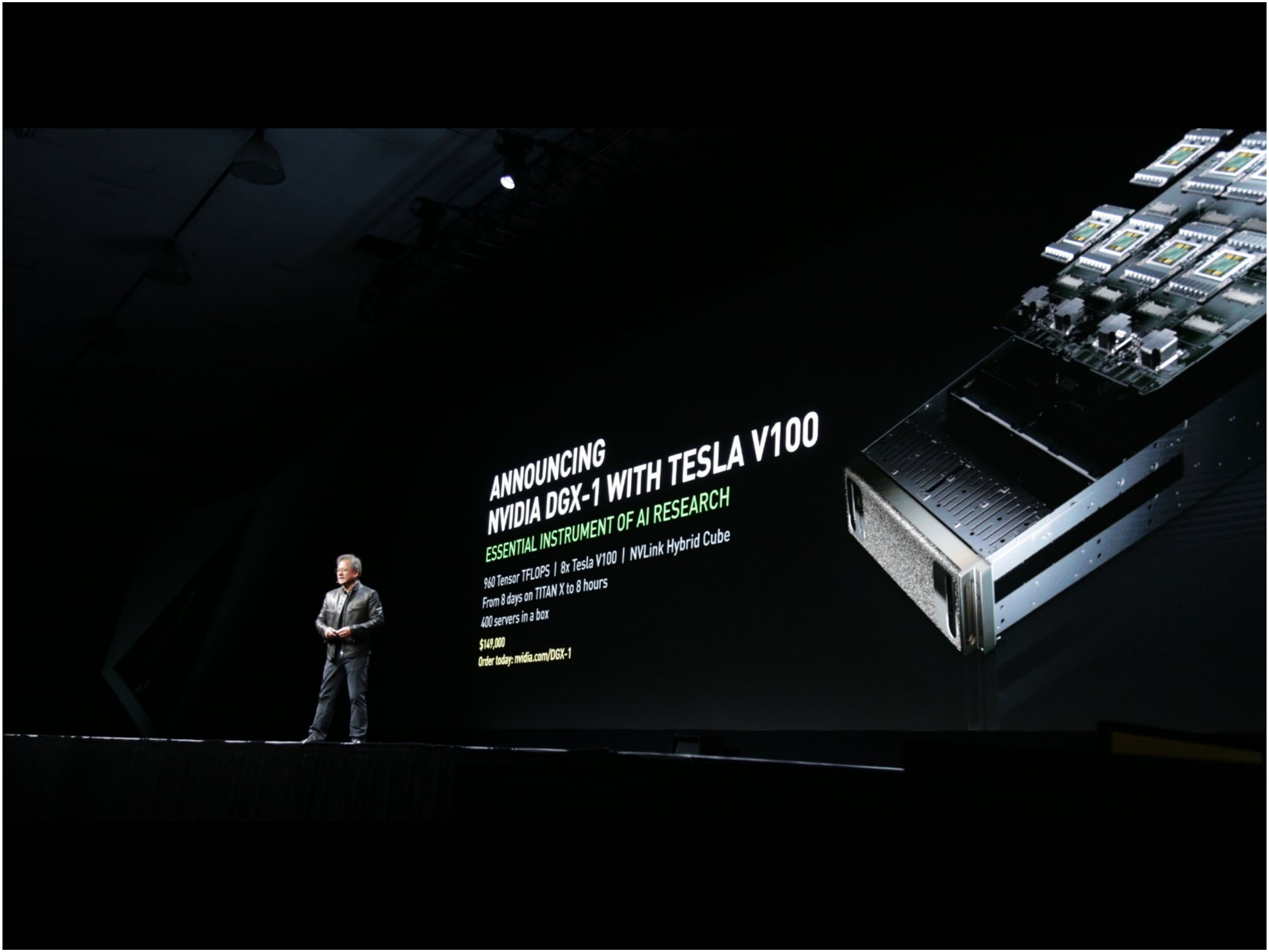
Source: Google

Note: 2015 data does not incorporate data from Q4

Bloomberg

By 2020, the market for machine learning will reach $40 billion, according to market research firm IDC.

Deep Learning market is projected to be ~$5B by 2020



ANNOUNCING
NVIDIA DGX-1 WITH TESLA V100

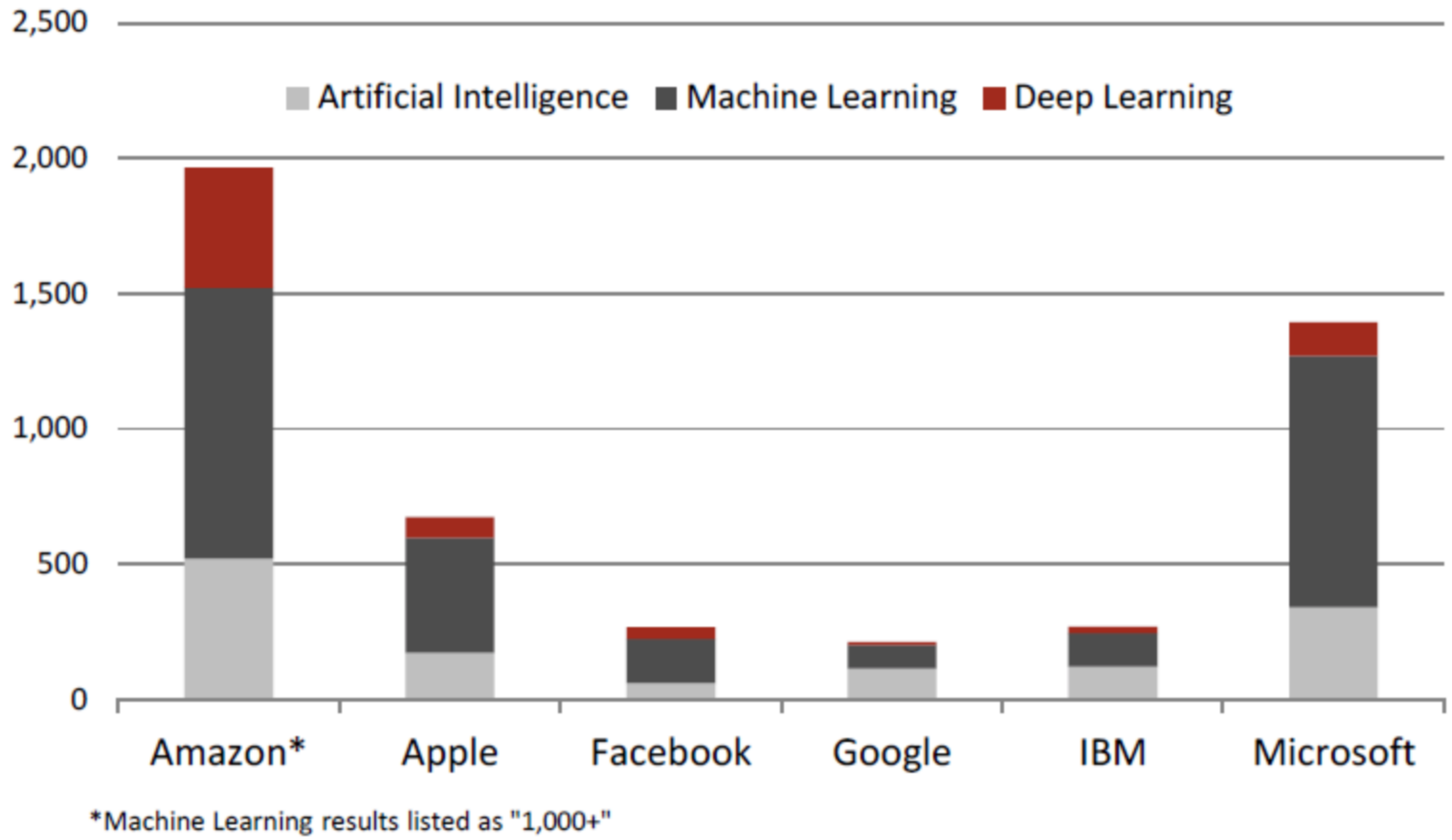
ESSENTIAL INSTRUMENT OF AI RESEARCH

960 Tensor TFLOPS | 8x Tesla V100 | NVLink Hybrid Cube
From 8 days on TITAN X to 8 hours
400 servers in a box

$149,000
Order today: nvidia.com/DGX-1

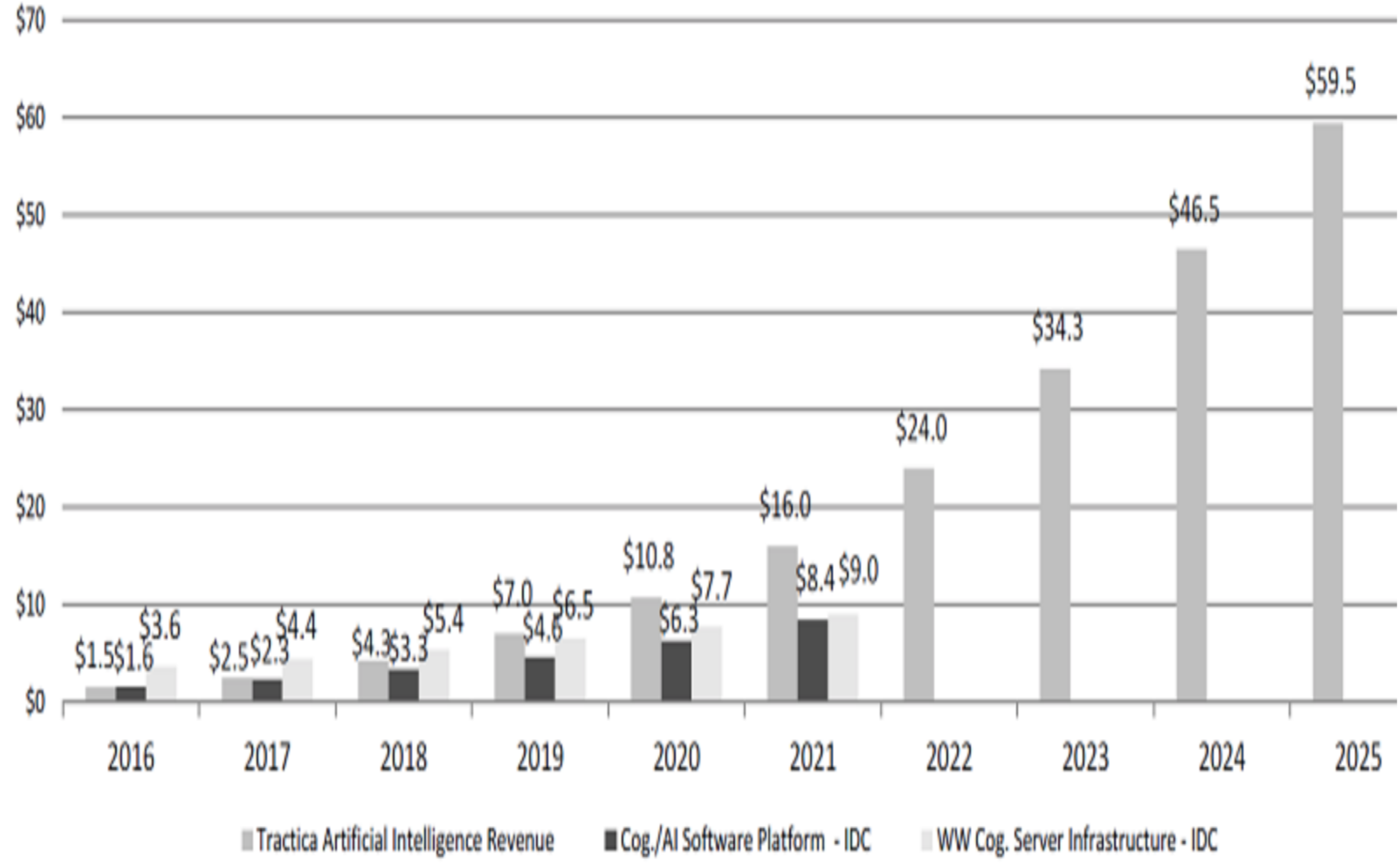
intel
AI AND MORE ON IA
Dr. Rajeeb Hazra
Vice President, Data Center Group
General Manager, Enterprise and Government Group
ISC High Performance
The HPC Event
ISC High Performance
The HPC Event
ISC High Performance

Exhibit 23: Monster.com Postings by Company, Search Terms: Artificial Intelligence, Machine Learning, and Deep Learning



Source: monster.com

Exhibit 8: Artificial Intelligence Industry Forecasts ($B)



Source: Jefferies, Gartner, IDC, Tractica

Markets are Developing at Different Rates ~2020

- HPC (Simulation)→ ~$30B @ 5.45%
- Data Analysis → ~$200B @ 11.7%
- Deep Learning → ~$5B @ 65%

- DL > HPC in 2024
- DL > DA in 2030

Big Picture

- Mix of applications is changing
- HPC “Simulation”, “Big” Data Analytics, Machine Learning “AI”
- Many projects are combining all three modalities
  - Cancer
  - Cosmology
  - Materials Design
  - Climate
  - Drug Design

Deep Learning in Climate Science

- Statistical Downscaling
- Subgrid Scale Physics
- Direct Estimate of Climate Statistics
- Ensemble Selection
- Dipole/Antipode Detection

Review
Chinese Science Bulletin
Ecology
December 2010 Vol.55 No.34: 3853–3863
doi: 10.1007/s11434-010-4183-3

Application of artificial neural networks in global climate change and ecological research: An overview

LIU ZeLin[1], PENG ChangHui[1,2*], XIANG WenHua[1], TIAN DaLun[1], DENG XiangWen[1] & ZHAO MeiFang[1]

[1] *College of Life Science and Tec*
[2] *Institute of Environment Scienc*
Montreal (QC) H3C 3P8, Cana

Received May 7, 2010; accepted

Fields that employ artificial n
development of computer tec
ers in light of increasing con
events, and atmospheric circ
rameters require typically hi
ual experience. ANNs, howe
search favoring the effects o
remain a better choice than n
study of global climate chan
true for situations in which n
ANNs will be widely adopte

global change, ecology, arti

Citation: Liu Z L, Peng C H, Xi
Sci Bull, 2010, 55: 385

In recent years, global cli
of attention for governm
large. Related research in
ducted and remarkable
Global climate change is
ing, frequent El Niño and
circulation anomalies. Th
increased by approximate
and is projected to contin
tion, global atmospheric
nearly 38% since the pre-i
to surpass 700 µmol mol

*Corresponding author (email: peng.

© Science China Press and Sprin

Neural Network Modeling in Climate Change Studies 12

Antonello Pasini

12.1 Introduction

At present, climate change is a "hot topic", not only in scientific analyses and papers by researchers, but also in wider discussions among economists and policy-makers.

In whatever area you are, the role of modeling appears crucial in order to understand the behavior of the climate system and to grasp its complexity. Furthermore, once validated on the past, a model represents the only chance to make projections about the future behavior of the climate system.

In this framework, AI methods (more specifically, neural networks – NNs) have recently shown their usefulness in modeling studies dealing with the climate system. Thus, the aim of this paper is to review and discuss the applications of neural network modeling
to climate
at least tw
"complem
via Globa
stantial na
the secon
modeling,
of the clim
driven wa
Thus,
ies about
activity t
papers du

thoroughly applied. Then, considerations about the application of dynamical modeling to a complex system like climate will lead us to recognize some weaknesses in the simulated reconstruction of the system itself. In particular, as we will see, this can lead us to criticize this kind of modeling (when applied to complex systems) and to not rely on its results.

In this unsatisfactory situation, a more phenomenological approach to the analysis of climate behavior can be applied. In Section 12.3 we will initially show how some modelers used neural network modeling for studying and forecasting the occurrence of specific phenomena, like El Niño. Then, in what follows, a comprehensive analysis of the influence of natural and anthropogenic forcings on global and regional temperature behaviors will be performed by NNs, shedding light on the most important forcings that drove the observed trends in the past.

Antonello P
CNR – Inst
29.300, I-0
Phone: + 3
email: pasi

S. E. Haupt
© Springer

Theor Appl Climatol (2011) 103:103–107
DOI 10.1007/s00704-010-0285-8

ORIGINAL PAPER

Detecting human influence on climate using neural networks based Granger causality

A. Attanasio · U. Triacca

Received: 17 August 2009 / Accepted: 30 March 2010 / Published online: 25 April 2010
© Springer-Verlag 2010

Abstract In this note we observe that a problem of linear approach to Granger causality testing between CO_2 and global temperature is that such tests can have low power. The probability to reject the null hypothesis of non-causality when it is false is low. Regarding non-linear Granger causality, based on multi-layer feed-forward neural network, the analysis provides evidence of significant unidirectional Granger causality from CO_2 to global temperature.

1 Introduction

There is little doubt that much of the Earth has been undergoing a pronounced warming since about the start of the twentieth century. An important question is: Has this global warming been primarily due to natural fluctuations or anthropogenic influences? Among a number of man-made 'greenhouse' gases which absorb and emit thermal radiation, such as nitrous oxide and methane, CO_2 is widely believed to be the major 'greenhouse' gas which has the greatest influence on the global climate. Consequently, the relationships between global temperature and CO_2 concentration have been the subject of intense research in the last two decades.

In particular Triacca (see Triacca 2005) finds that there is no detectable Granger causality (see Granger 1969) from radiative forcing due to CO_2 concentration to global surface temperature anomalies. In other terms, past observations of radiative forcing due to CO_2 do not significantly improve the predictability of current temperature. In this note we show that this (negative) result is due to the inappropriateness of the linear version of the Granger non-causality test utilized by the author. The problem of linear approach to causality testing is that such tests can have low power detecting certain kinds of non-linear causal relations. Regarding non-linear Granger causality based on multi-layer feed-forward (MLF) neural network between CO_2 and global temperature, the analysis provides a quite different story. It provides evidence of significant unidirectional non-linear Granger causality from CO_2 to global temperature.

The paper is organized as follows: A short review of neural networks considering their usage in this work is given in the next section; The results of non-linear Granger causality analysis are presented in Section 3; Some concluding remarks are given in Section 4.

2 Methodology and data

We are interested to test the null hypothesis that CO_2 concentration does not Granger cause global surface temperature. The linear framework is too simple to catch the complex non-linear behavior of the temperature with respect CO_2 concentration. To overcome this problem we use a non-linear out-of-sample analysis of the dynamic relationship between these variables based on artificial neural network models.

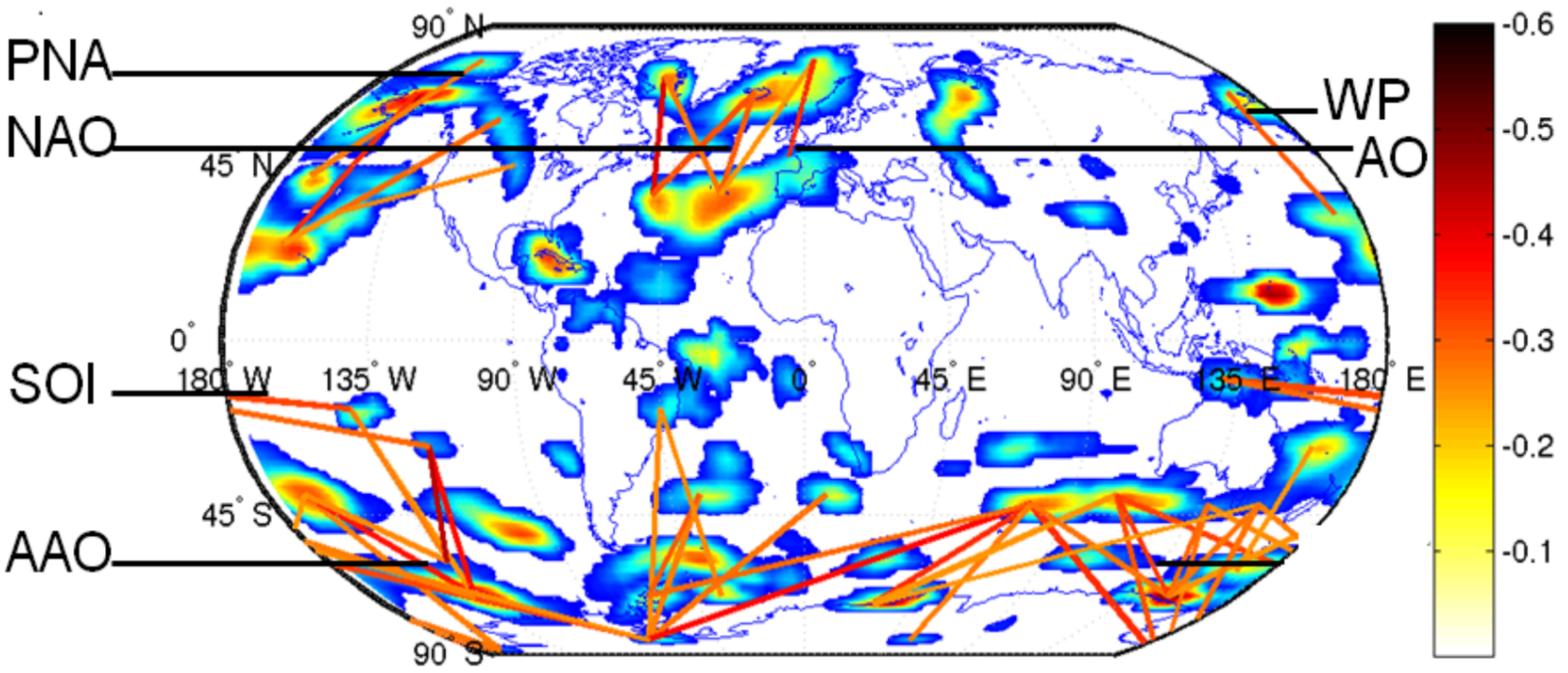
Climate change studies represent a field in which neural networks can be applied successfully (see Pasini

A. Attanasio · U. Triacca (✉)
Università di L'Aquila, L'Aquila, Italy
e-mail: umberto.triacca@ec.univaq.it

A. Attanasio
e-mail: alessandro_attanasio@yahoo.it

Springer

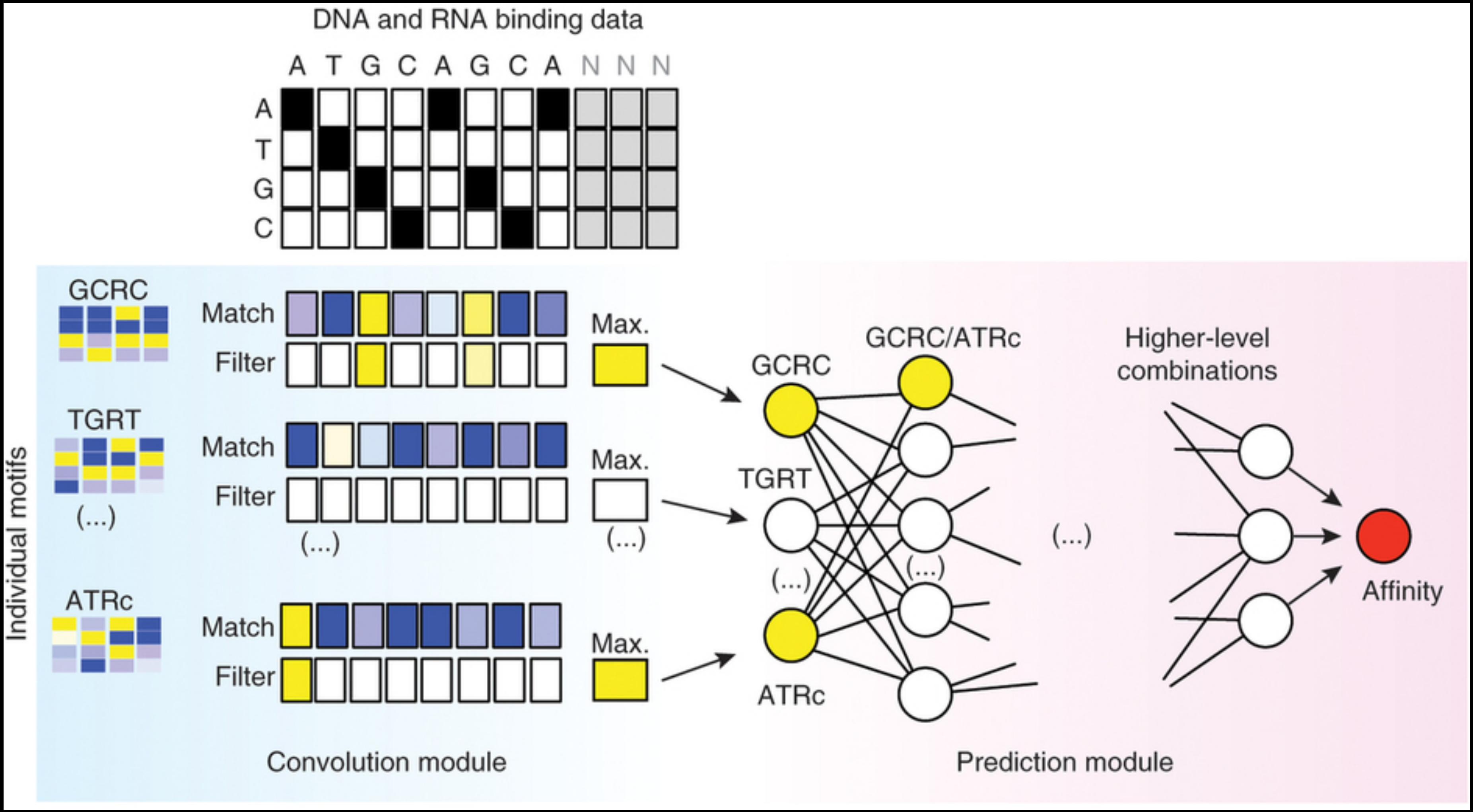
Automatic Discovery of Dipoles



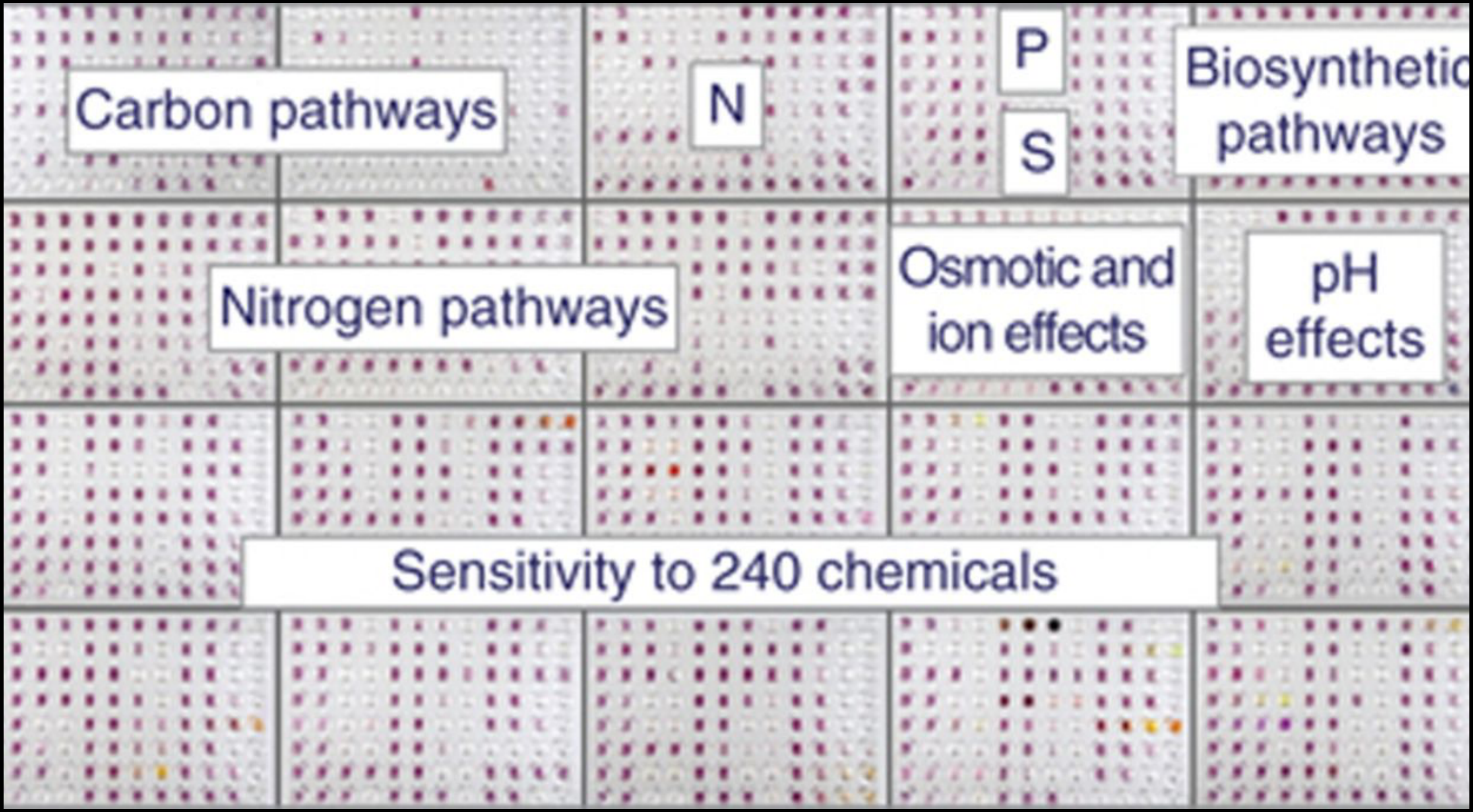
Credit: V. Kumar

- Detection of Global Dipole Structures
  - Most known dipoles discovered
  - Some `new' dipoles: Previously unknown phenomenon?
  - A new dipole near Australia [Liess et al., J Clim'14]

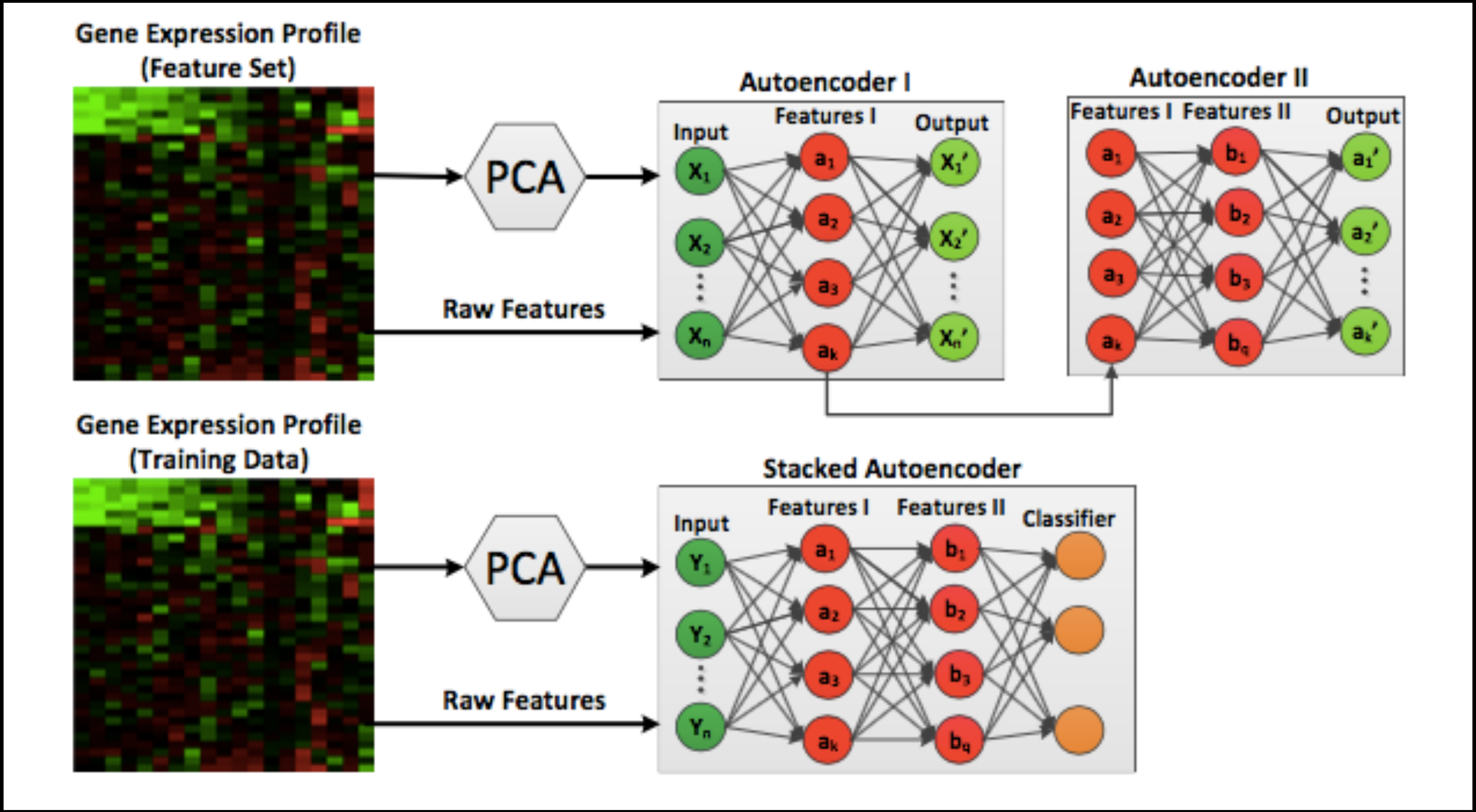
Deep Learning in Genomics



Predicting Microbial Phenotypes



Classification of Tumors



Using deep learning to enhance cancer diagnosis and classification, ICML2013

High Throughput Drug Screening

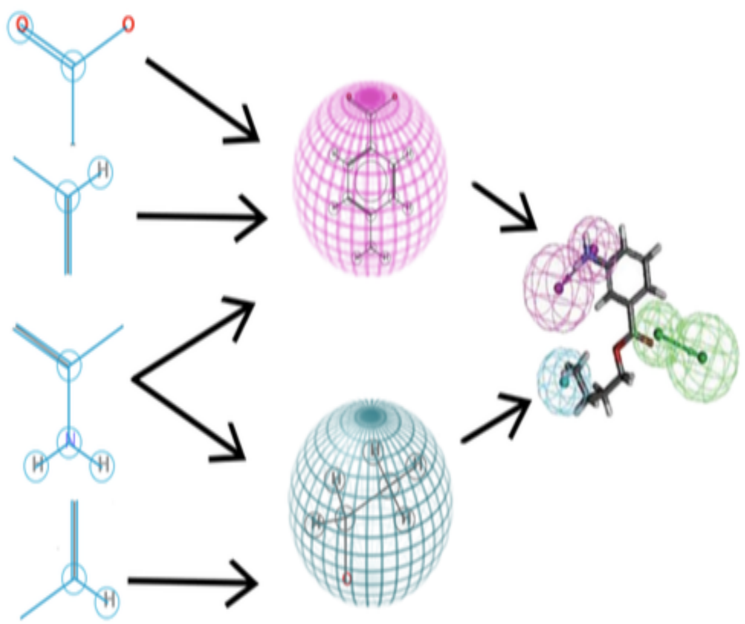
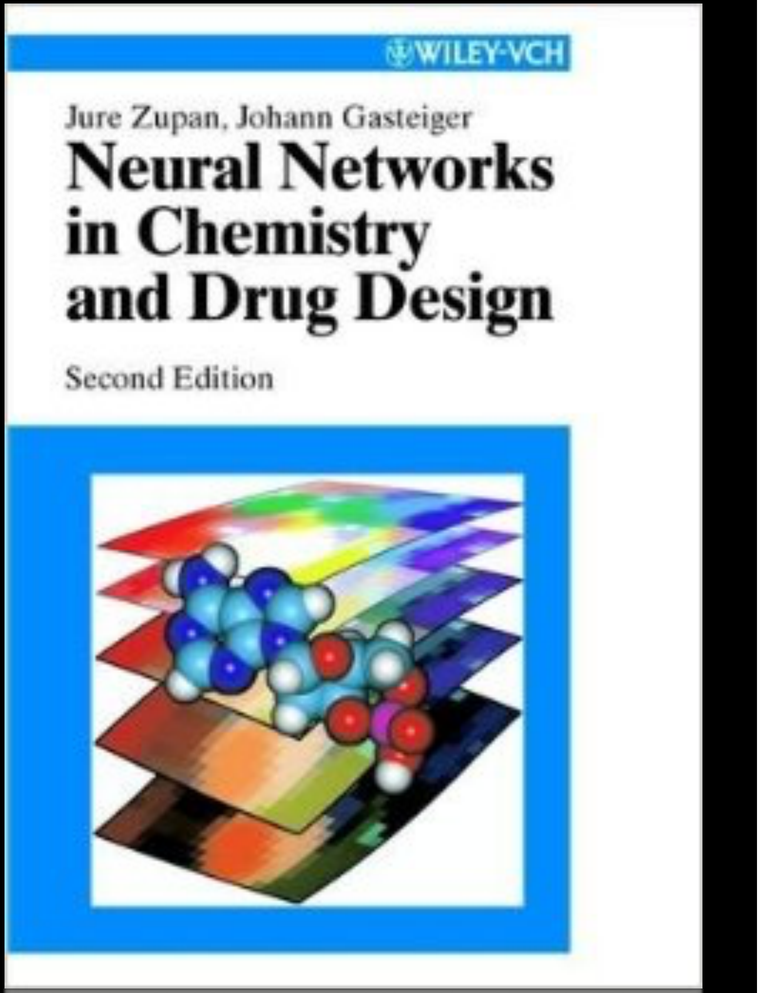


Figure 1: Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.



WILEY-VCH

Jure Zupan, Johann Gasteiger

Neural Networks in Chemistry and Drug Design

Second Edition

Deep Learning as an Opportunity in Virtual Screening, NIPS2014

Deep Networks Screen Drugs

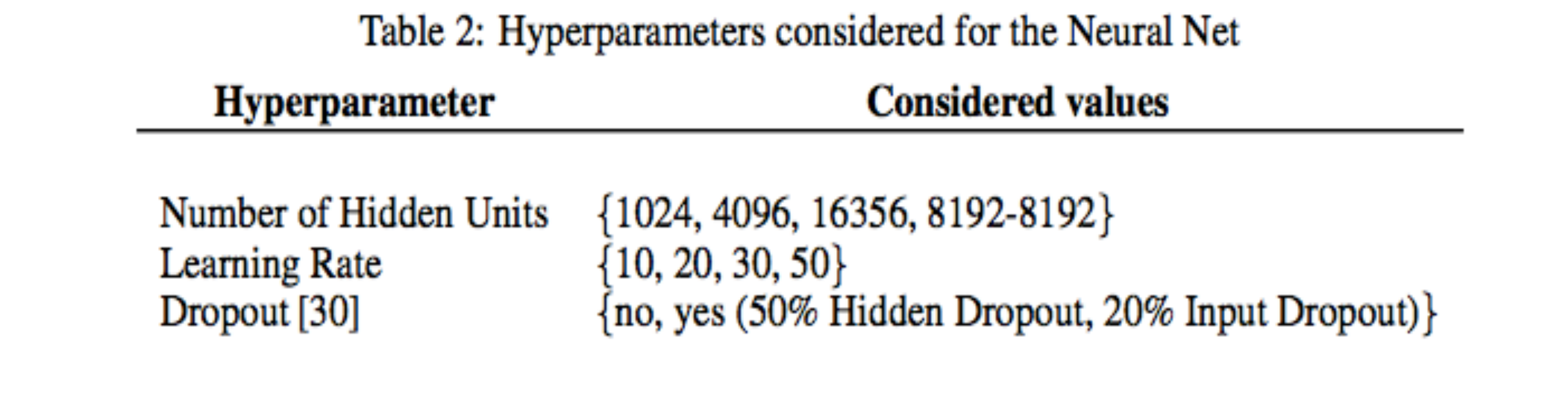
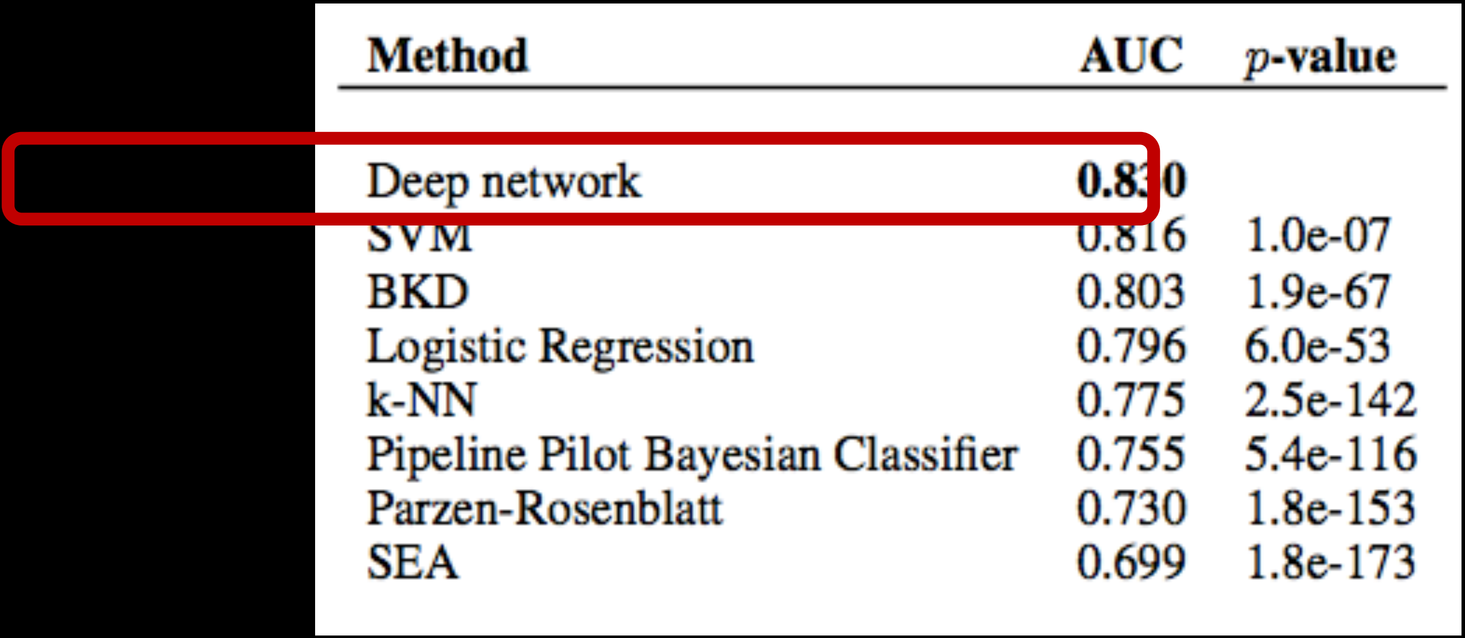


Table 2: Hyperparameters considered for the Neural Net

Hyperparameter	Considered values
Number of Hidden Units	{1024, 4096, 16356, 8192-8192}
Learning Rate	{10, 20, 30, 50}
Dropout [30]	{no, yes (50% Hidden Dropout, 20% Input Dropout)}



Method	**AUC**	***p*-value**
Deep network	**0.830**	
SVM	0.816	1.0e-07
BKD	0.803	1.9e-67
Logistic Regression	0.796	6.0e-53
k-NN	0.775	2.5e-142
Pipeline Pilot Bayesian Classifier	0.755	5.4e-116
Parzen-Rosenblatt	0.730	1.8e-153
SEA	0.699	1.8e-173

Deep Learning and Drug Discovery

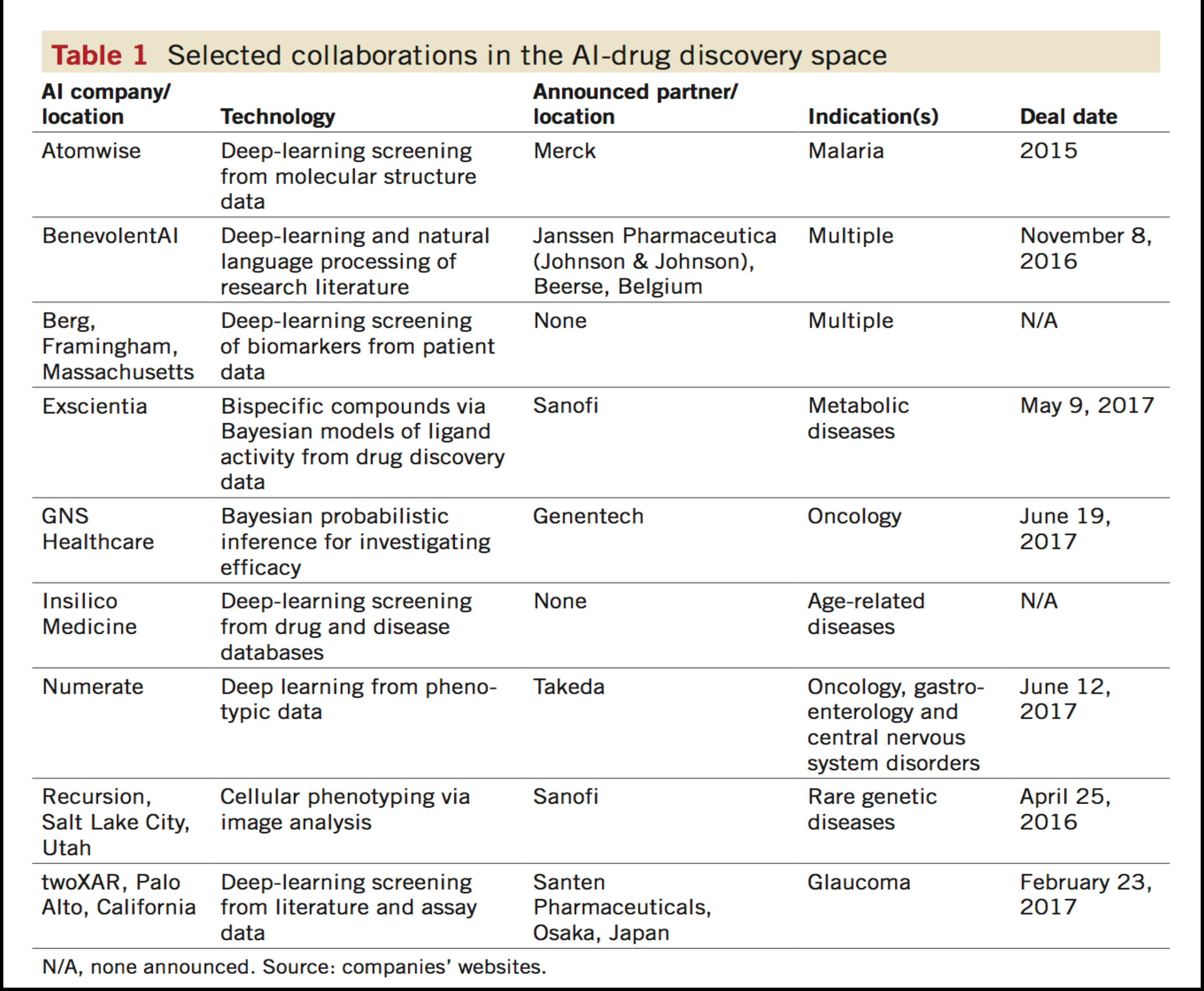
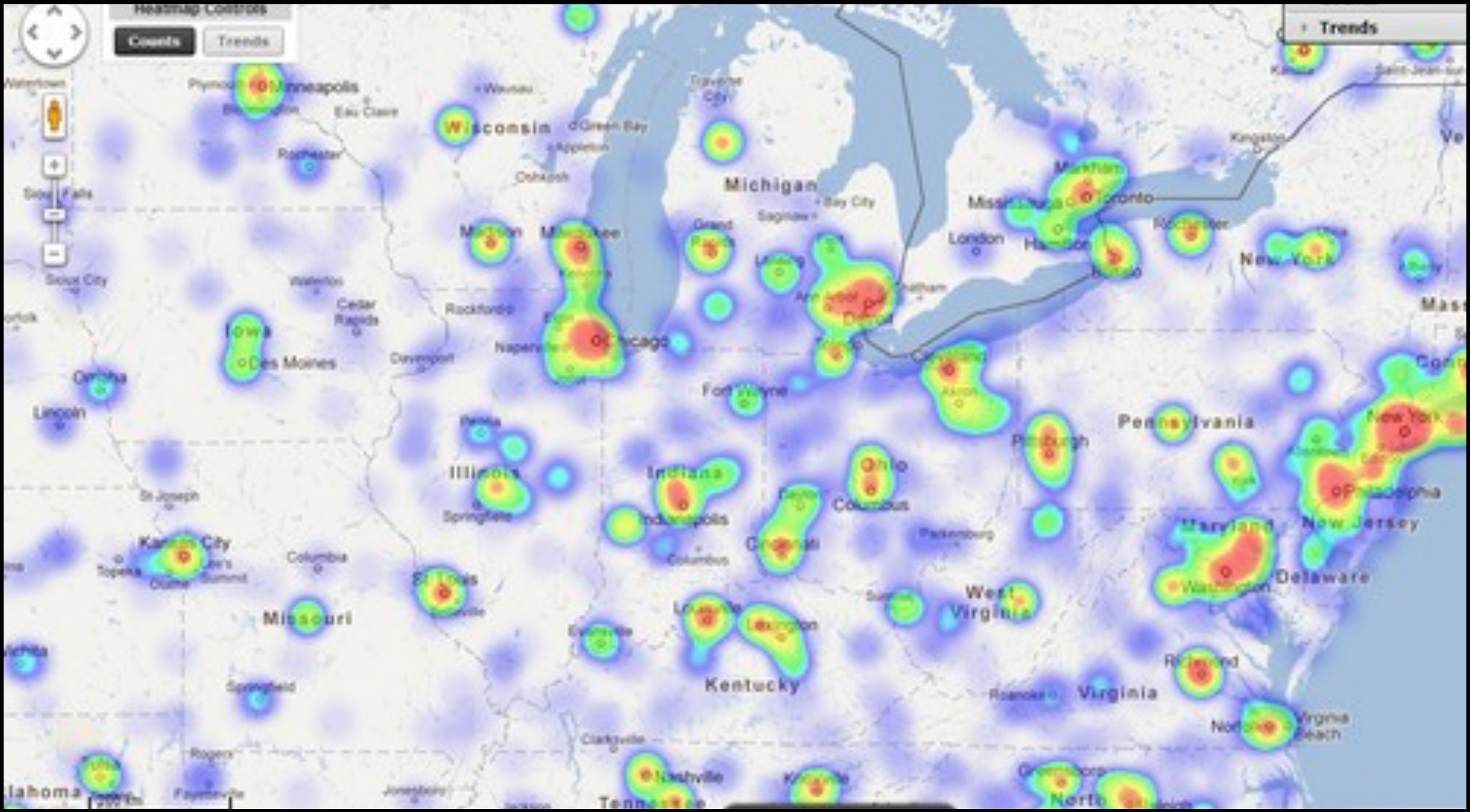


Table 1 Selected collaborations in the AI-drug discovery space

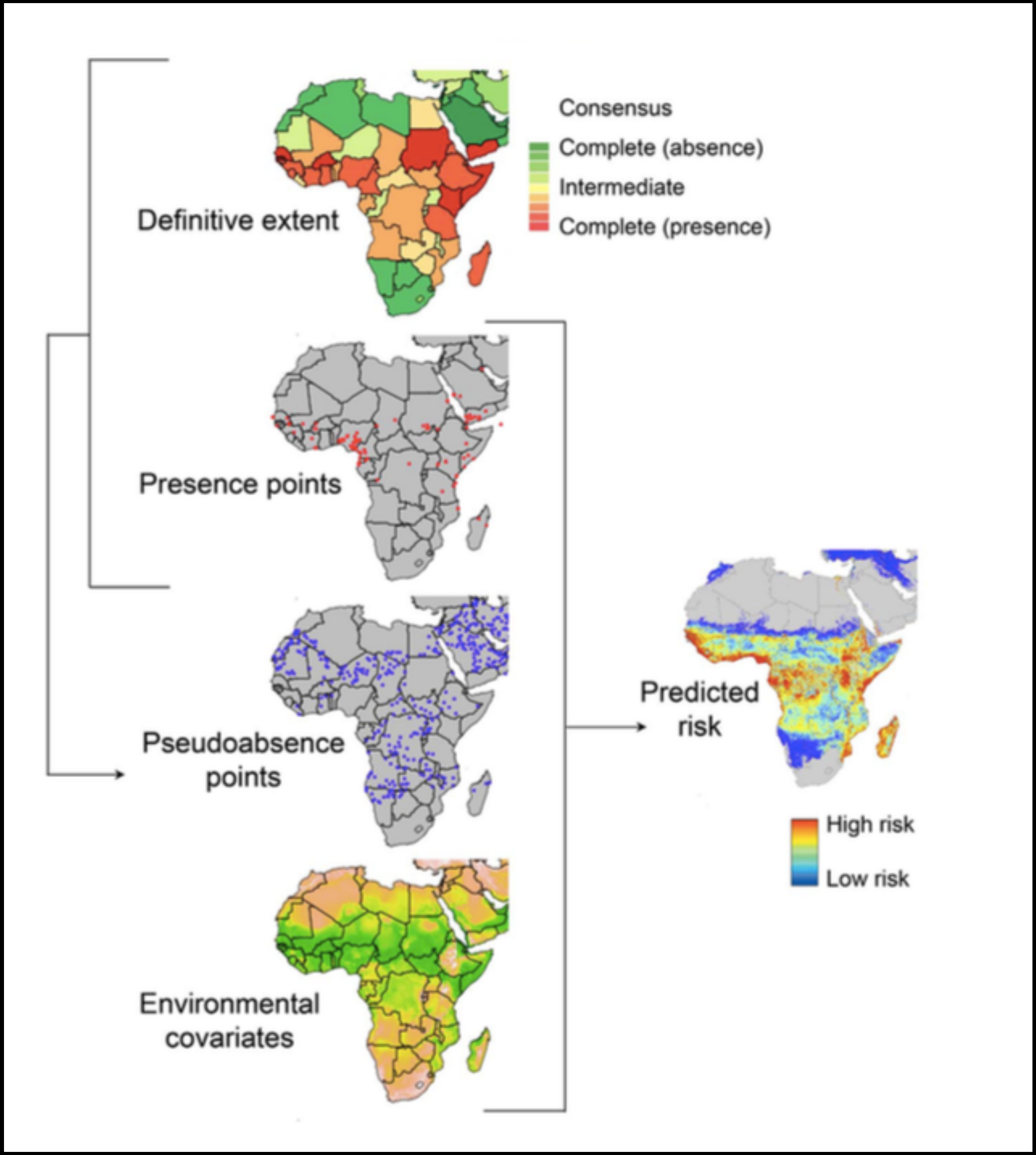
AI company/ location	Technology	Announced partner/ location	Indication(s)	Deal date
Atomwise	Deep-learning screening from molecular structure data	Merck	Malaria	2015
BenevolentAI	Deep-learning and natural language processing of research literature	Janssen Pharmaceutica (Johnson & Johnson), Beerse, Belgium	Multiple	November 8, 2016
Berg, Framingham, Massachusetts	Deep-learning screening of biomarkers from patient data	None	Multiple	N/A
Exscientia	Bispecific compounds via Bayesian models of ligand activity from drug discovery data	Sanofi	Metabolic diseases	May 9, 2017
GNS Healthcare	Bayesian probabilistic inference for investigating efficacy	Genentech	Oncology	June 19, 2017
Insilico Medicine	Deep-learning screening from drug and disease databases	None	Age-related diseases	N/A
Numerate	Deep learning from phenotypic data	Takeda	Oncology, gastroenterology and central nervous system disorders	June 12, 2017
Recursion, Salt Lake City, Utah	Cellular phenotyping via image analysis	Sanofi	Rare genetic diseases	April 25, 2016
twoXAR, Palo Alto, California	Deep-learning screening from literature and assay data	Santen Pharmaceuticals, Osaka, Japan	Glaucoma	February 23, 2017

N/A, none announced. Source: companies' websites.

Deep Learning In Disease Prediction



Learning Climate Disease Environment Associations



Big Data Opportunities for Global Infectious Disease Surveillance
Simon I. Hay, Dylan B. George, Catherine L. Moyes, John S. Brownstein

Neural Networks
in
Materials science

- Estimate Materials Properties from Composition Parameters
- Estimate Processing Parameters for Synthesis
- Materials Genome

JOURNAL OF MATERIALS SCIENCE: MATERIALS IN ELECTRONICS **16** (2005) 673–679

On the use of a neural network to characterize the plasma etching of SiON thin films

B. KIM
Department of Electronic Engineering, Bio Engineering Research Center, Sejong University, 98, Goonja-Dong, Kwangjin-Gu, Seoul, 143–747, Korea
E-mail: kbwhan@sejong.ac.kr

B. T. LEE
Department of Materials Science and Engineering, Chonnam National University, 300, Yongbong-Dong, Buk-Ku, Kwangju-Si, 500–757, Korea

K. K. LEE
Division of Micromechatronics, Korea Institute of Industrial Technology, Chunan, South Korea

Using a generalized regression neural network (GRNN), plasma etching of oxynitride thin
films was modeled. The et
experiment. A genetic alg
optimizing multiparamete
the constructed etch rate r
prediction performance. 3
mechanisms while validat
and chemical effects, both
source power affected sig
or C_2F_6 flow rate. For pres
chemical etching. The con
chemical etching or polym

1. Introduction

In manufacturing optical devic
(SiON) film is a promising ma
pability to achieve a higher ref
between the core and cladding la
tures attractive for manufacturin
vices include low density of surf
tric permittivity, and controllabi
terms of [O]/[N] ratio [4]. Mos
on studying deposition characte
fractive index or electrical prop
devices. Few studies have been
characteristics of SiON films. Pl
means to form fine patterns in ma
devices. In plasma etching, man
are typically included and thei
siderably affect etching characte
experimental budget, it is not po
rameter effects under various pl
limitation can be overcome by
puter prediction model. Despite
on plasma dynamics, first princ
ject to many assumptions, resul
between the predictions and actu
other alternative is to use a neu
in conjunction with a statistical

0957–4522 © *2005 Springer Scien*

J Mater Sci
DOI 10.1007/s10853-011-5868-y

A neural network approach for the prediction of the refractive index based on experimental data

Alex Alexandridis · Eva Chondrodima · Konstantinos Moutzouris · Dimos Triantis

Received: 26 May 2011 / Accepted: 10 August 2011
© Springer Science+Business Media, LLC 2011

Abstract This article presents a systematic approach for correlating the refractive index of different material kinds and forms with experimentally measured inputs like wavelength, temperature, and concentration. The correlation is accomplished using neural network models, which can deal effectively with the nonlinear nature of the problem without requiring a predefined form of equation, while taking into account all the parameters affecting the refractive index. The proposed methodology employs the
powerful radial basi
the neural network
using an innovative a
increased prediction
to two cases, involv
index of semiconduc
water mixture and th
predictions are accur
of decimal places as
with other neural net
empirical forms like
superiority of the pr

fundamental physical property of substance related not only to its optical, but also electrical, magnetic, thermal, and mechanical properties [1–4]. In general, *n* depends on light wavelength and temperature, effects commonly referred to as chromatic and temperature dispersion, respectively. However, in many situations there exist numerous additional parameters influencing the refractive index, ranging from doping level and composition in amorphous materials and semiconductor or dielectric

Introduction

The refractive index
of the velocity of ligh
light in the consider

A. Alexandridis (✉) · E
D. Triantis
Laboratory of Electric P
of Electronics, Technolo
Agiou Spiridonos, 1221
e-mail: alexx@teiath.gr

Published online: 24 Au

Bull. Mater. Sci., Vol. 36, No. 7, December 2013, pp. 1307–1313. © Indian Academy of Sciences.

Quantitative structure-property relationships of electroluminescent materials: Artificial neural networks and support vector machines to predict electroluminescence of organic molecules

ALANA FERNANDES GOLIN and RICARDO STEFANI*
Laboratório de Estudos de Materiais (LEMAT), Instituto de Ciências Exatas e da Terra, Av. Governador Jaime Campos 6390, Campus Universitário do Araguaia, Universidade Federal de Mato Grosso, 78600-00 Barra do Garças – MT. Brazil

MS received 13 February 2012; revised 13 December 2012

Abstract. Electroluminescent compounds are extensively used as materials for application in OLED. In order to understand the chemical features related to electroluminescence of such compounds, QSPR study based on neural network model and support vector machine was developed on a series of organic compounds commonly used in OLED development. Radial-basis function-SVM model was able to predict the electroluminescence with good accuracy ($R = 0·90$). Moreover, RMSE of support vector machine model is approximately half of RMSE observed for artificial neural networks model, which is significant from the point of view of model precision, as the dataset is very small. Thus, support vector machine is a good method to build QSPR models to predict the electroluminescence of materials when applied to small datasets. It was observed that descriptors related to chemical bonding and electronic structure are highly correlated with electroluminescence properties. The obtained results can help in understating the structural features related to the electroluminescence, and supporting the development of new electroluminescent materials.

Keywords. QSPR; neural networks; SVM; electroluminescence; OLED; organic materials.

1. Introduction

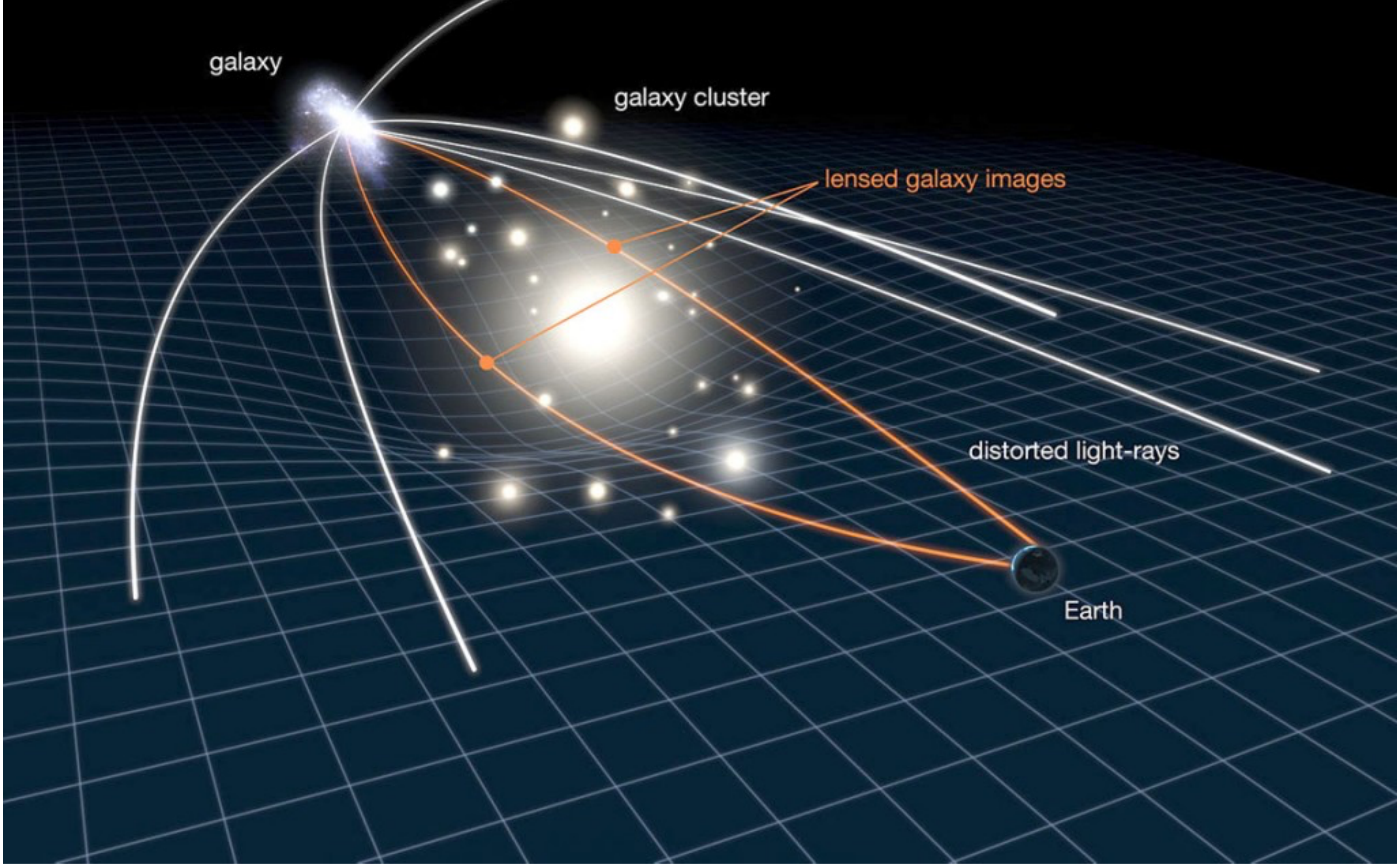
Electroluminescent materials (EL) are among the most promising modern materials with a wide range of technology applications (Xue and Luo 2003; So *et al* 2009). One of the most promising EL applications, is the design and fabrication of organic light-emitting diodes (OLEDs) (Akcelrud 2003). OLEDs have demonstrated manufacturing and market potential in small and medium device applications. Thus, OLED can become one of the mainstream display technologies, competing directly with LCD (liquid crystal display) technology (Wen *et al* 2005). For high-quality OLED displays, highly efficient and low-cost electroluminescent materials are of great importance, since, to gain market share over LCD displays, OLED devices need to be efficient and to have low prices to the final costumer. Many pyran-containing, polyaromatic hydrocarbons (PAH) and porphyrin type compounds are used in OLED fabrication and these compounds may be polymeric itself or used as a dopant to allow thinfilms to become electroluminescent (Mi *et al* 2002). Understanding of the physical and chemical features related to the electroluminescence of such materials, can help in the design and development of new chemical compounds with improved electroluminescence features. In order to develop new organic compounds that can be used in OLED applications, computational methods, such as quantitative–structure properties relationships (QSPR) have emerged as a fast and reliable method to predict and study physical–chemical properties of materials.

Quantitative–structure properties relationships (QSPR) models can be used to predict with good accuracy, key physical and chemical features from chemical compounds. QSPR methods are based on the existing correlation between groups of mathematical values (descriptors), representing certain features of a chemical structure and a target chemical property. The advantage of QSPR model is that it is based solely on the knowledge of chemical structure and it requires no additional experimental data and once the co-rrelation is established, it can be used for the prediction of properties of new compounds that have not been prepared (Yu *et al* 2008). Thus, QSPR models can be used to assist material design, since one can predict the properties of a certain material before its synthesis. As the development of new materials involves extensive experimental work, the ability to predict the properties of materials is of great value, because, it provides a guide to the development process and speeds up the development cycle, allowing time and reagent saving (Yu *et al* 2008). Thus, many research groups have been developing QSPR models in order to assist material discovery and design (Morril and Byrd 2008; Taherpour 2009; Fourches *et al* 2010; Yu 2010). The advantage of using QSPR models over traditional computational methods is that description calculation is quite easy and requires little computation time.

*Author for correspondence (rstefani@ufmt.br)

1307

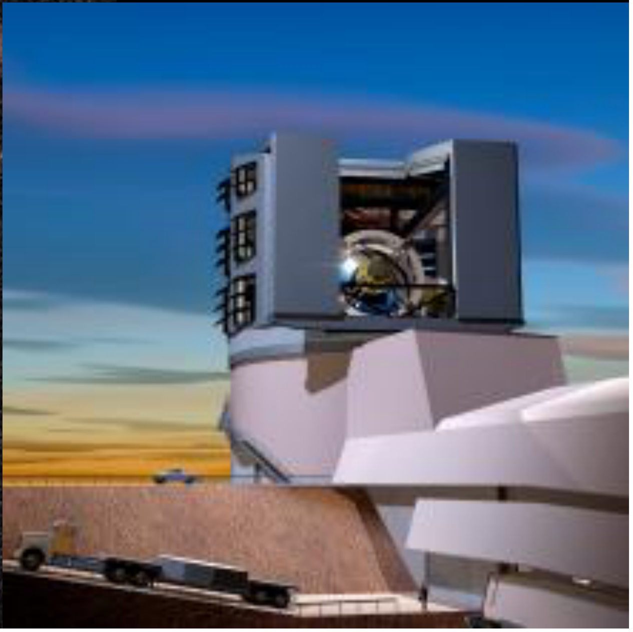
Searching For Lensed Galaxies



LSST

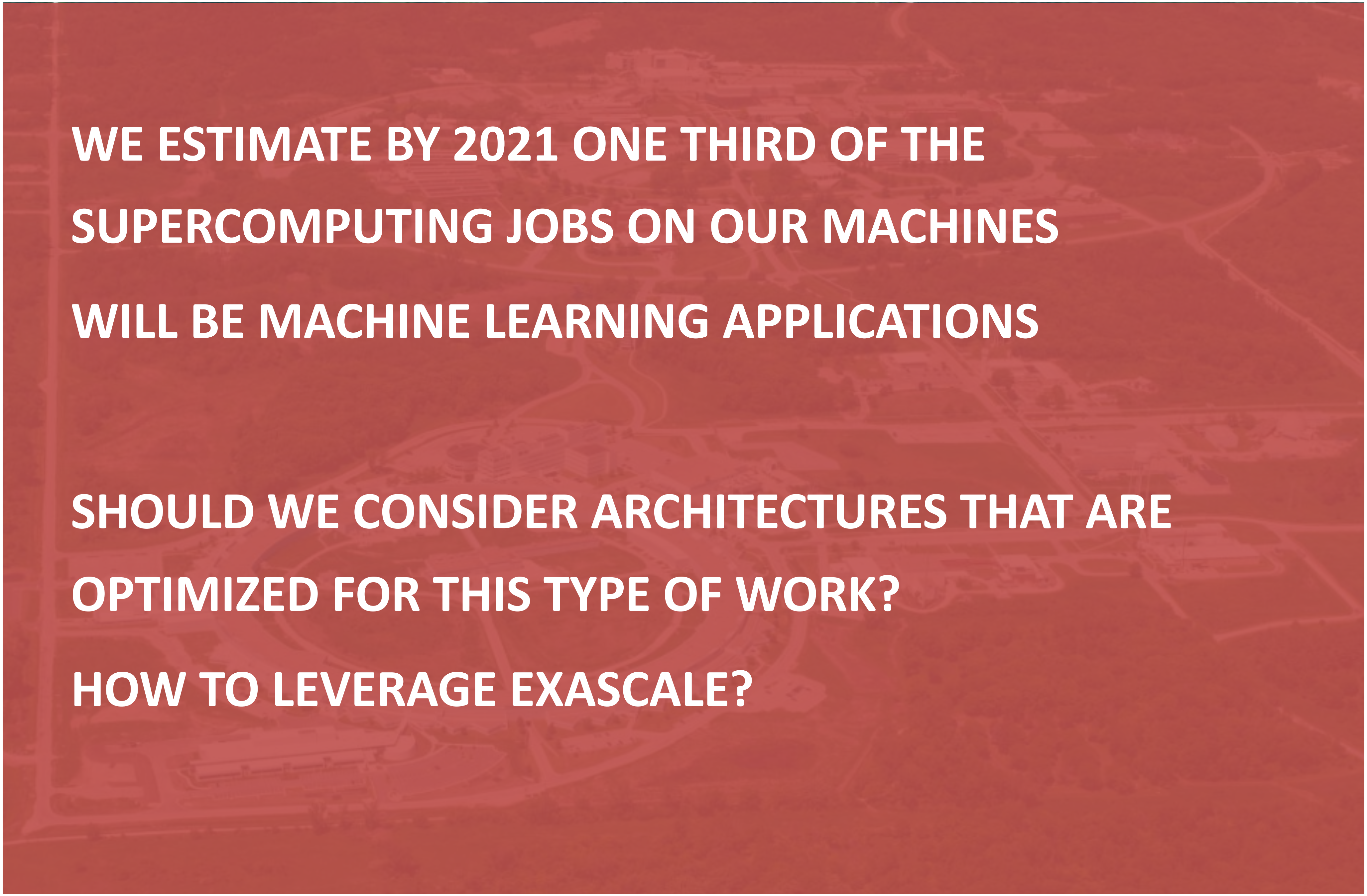
Large Synoptic Survey Telescope

15 TB/Night
Use CNN to find Gravitational Lenses



Deep Learning is becoming a major element of scientific computing applications

- Across the DOE lab system hundreds of examples are emerging
  - From fusion energy to precision medicine
  - Materials design
  - Fluid dynamics
  - Genomics
  - Structural engineering
  - Intelligent sensing
  - Etc.

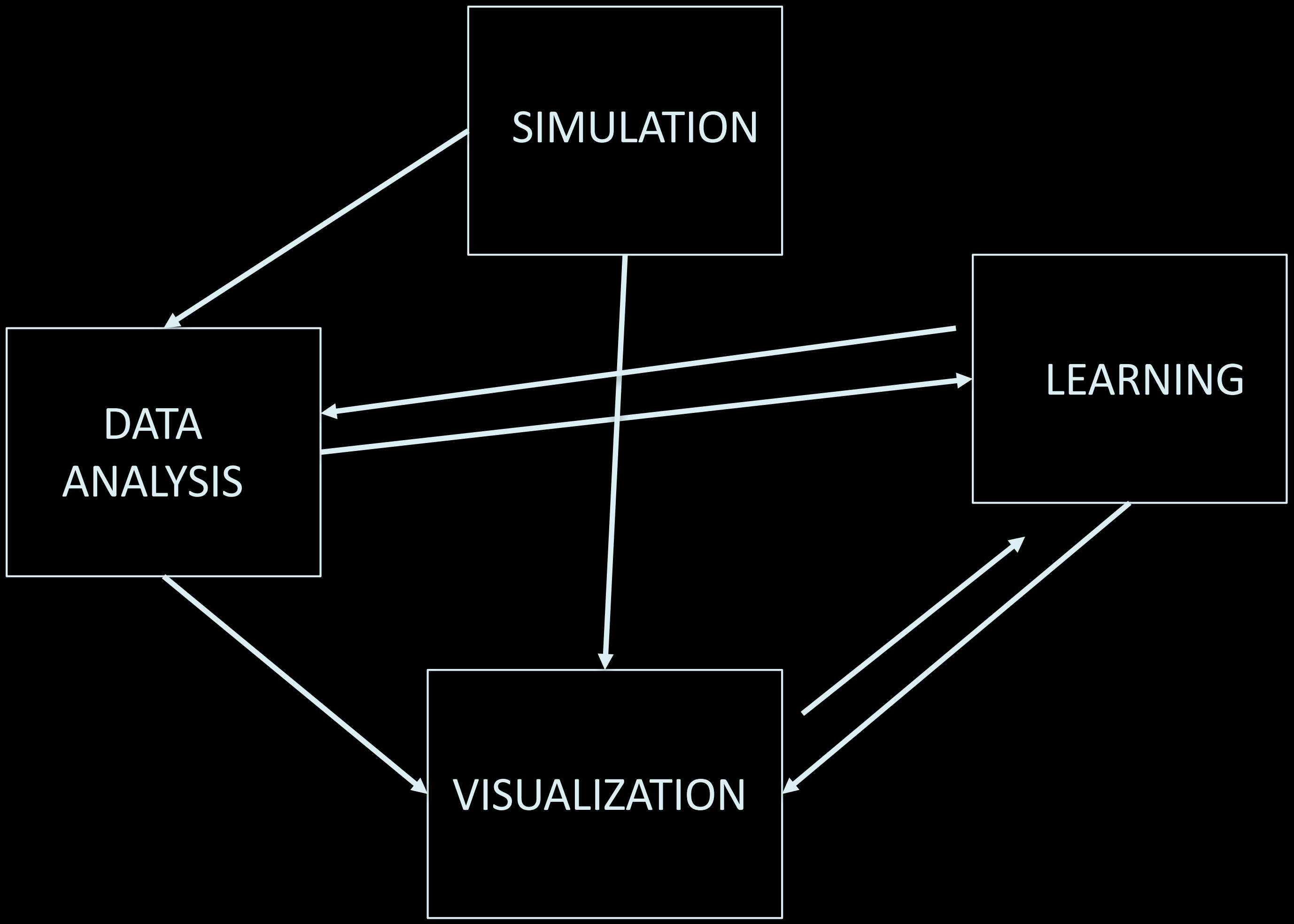


WE ESTIMATE BY 2021 ONE THIRD OF THE SUPERCOMPUTING JOBS ON OUR MACHINES WILL BE MACHINE LEARNING APPLICATIONS

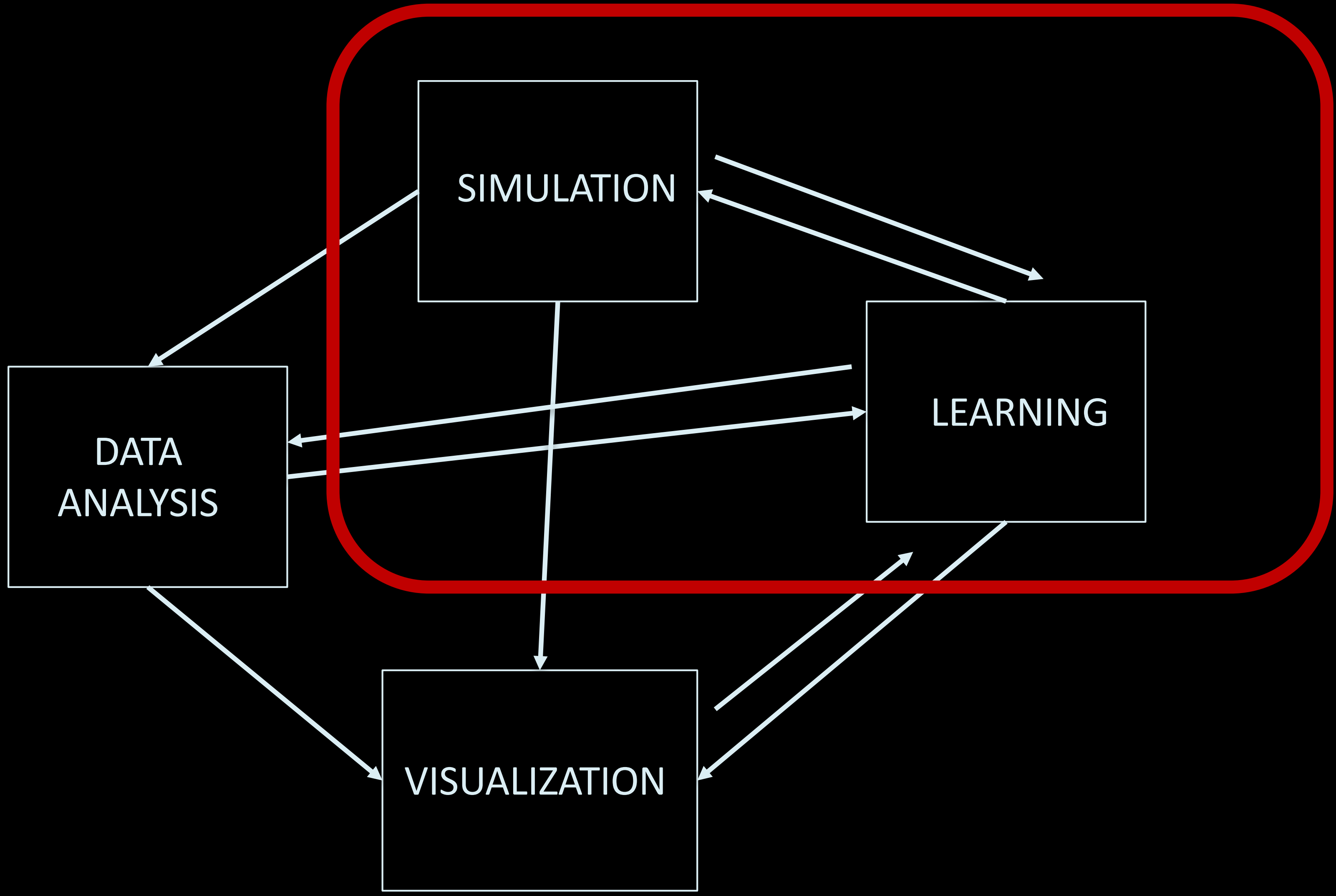
SHOULD WE CONSIDER ARCHITECTURES THAT ARE OPTIMIZED FOR THIS TYPE OF WORK?

HOW TO LEVERAGE EXASCALE?

The New HPC “Paradigm”



The New HPC "Paradigm"



The Critical Connections I

- Embedding Simulation into Deep Learning
  - Leveraging simulation to provide “hints” via the Teacher-Student paradigm for DNN
  - DNN invokes “Simulation Training” to augment training data or to provide supervised “labels” for generally unlabeled data
  - Simulations could be invoked millions of times during training runs
  - Training rate limited by simulation rates
  - Ex. Cancer Drug Resistance

Hybrid Models in Cancer

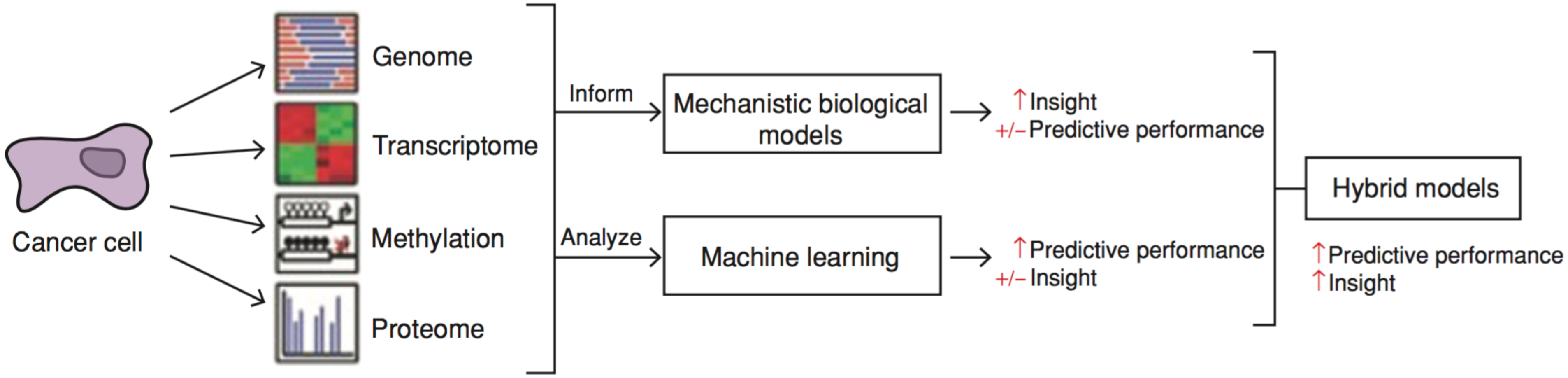
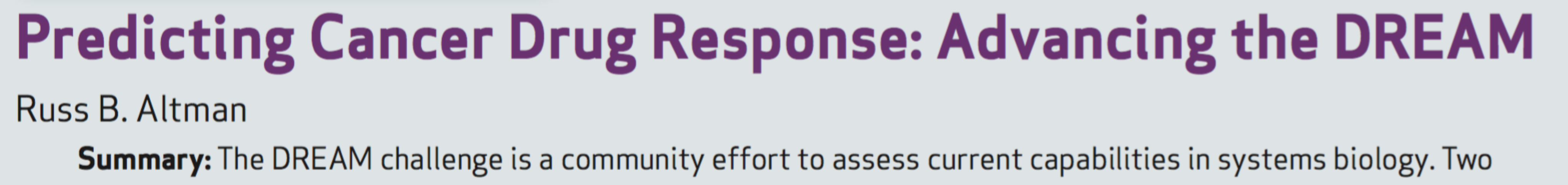


Figure 1. In two DREAM challenges, high throughput data characterizing cancer cells are used to build predictive models. Mechanistic models provide insight into the underlying biology, but do not take full advantage of the information within the data to achieve high performance. Machine learning methods are associative and extract maximum predictive value from the data, but do not always provide insight about mechanism. The future may bring hybrid models that combine the best of both approaches.

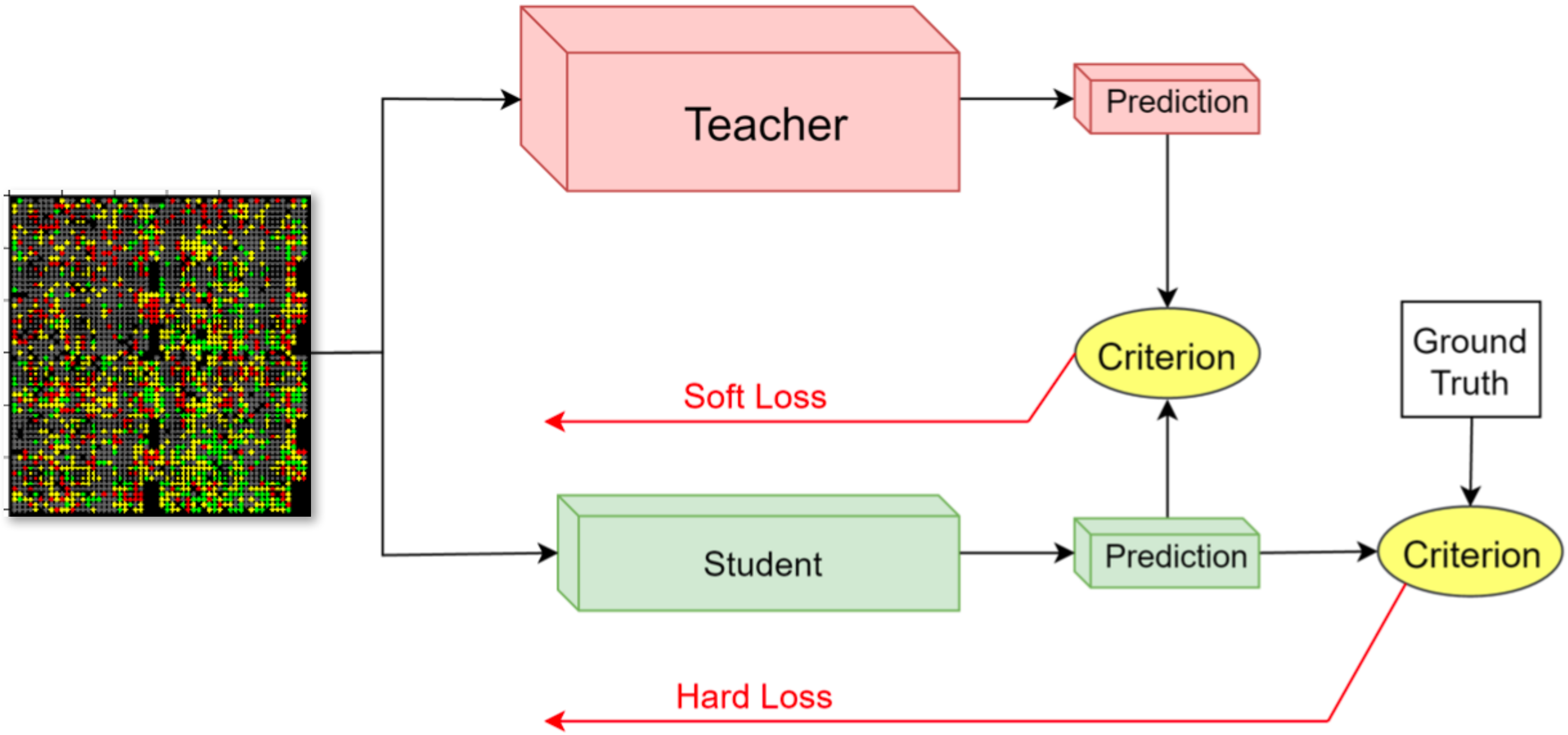


Predicting Cancer Drug Response: Advancing the DREAM

Russ B. Altman

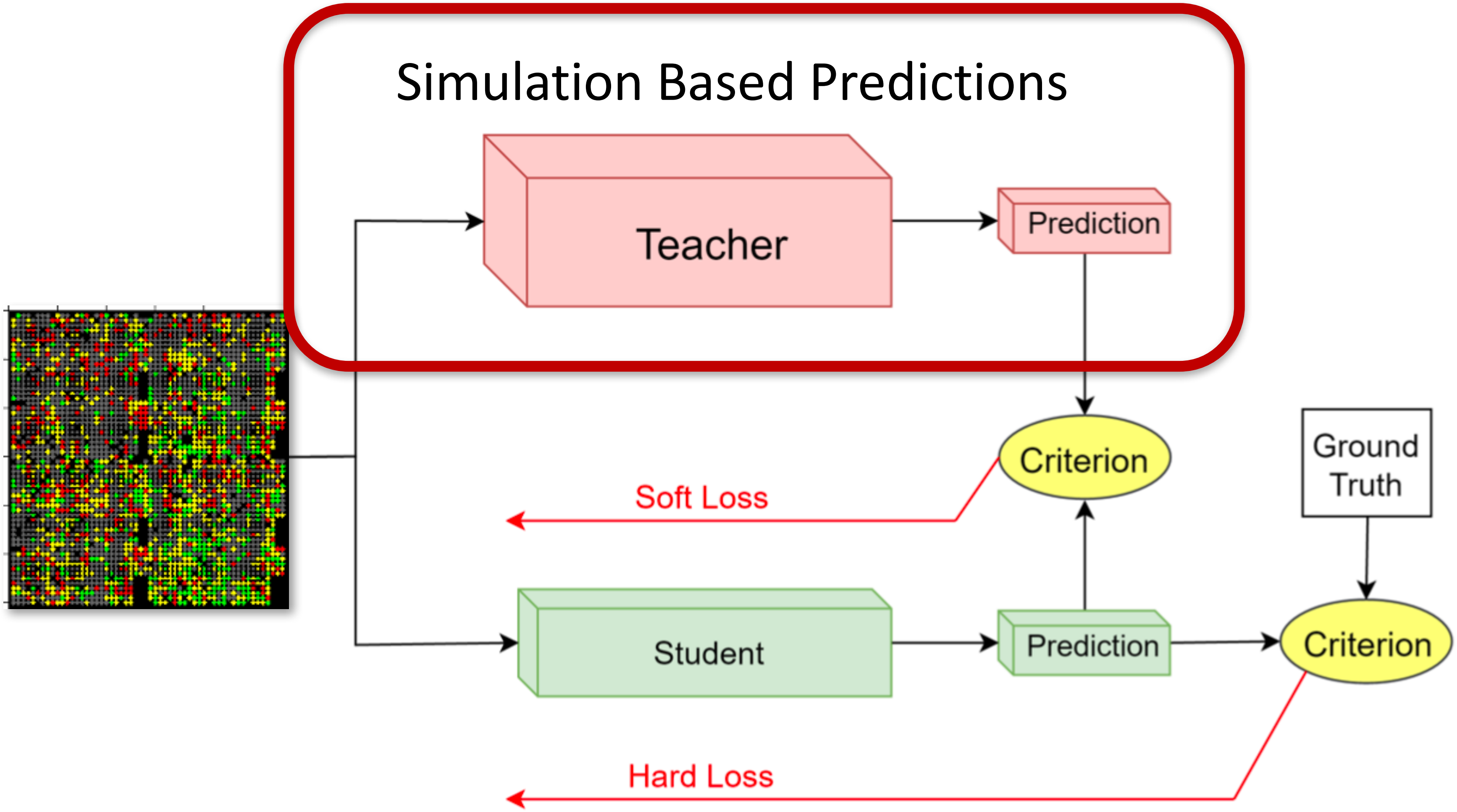
Summary: The DREAM challenge is a community effort to assess current capabilities in systems biology. Two

Teacher-Student Network Model



(a) Standard: student network learns from teacher guidance (soft loss) and ground truth (hard loss).

Teacher-Student Network Model



Simulation Based Predictions

(a) Standard: student network learns from teacher guidance (soft loss) and ground truth (hard loss).

Integrating ML and Simulation

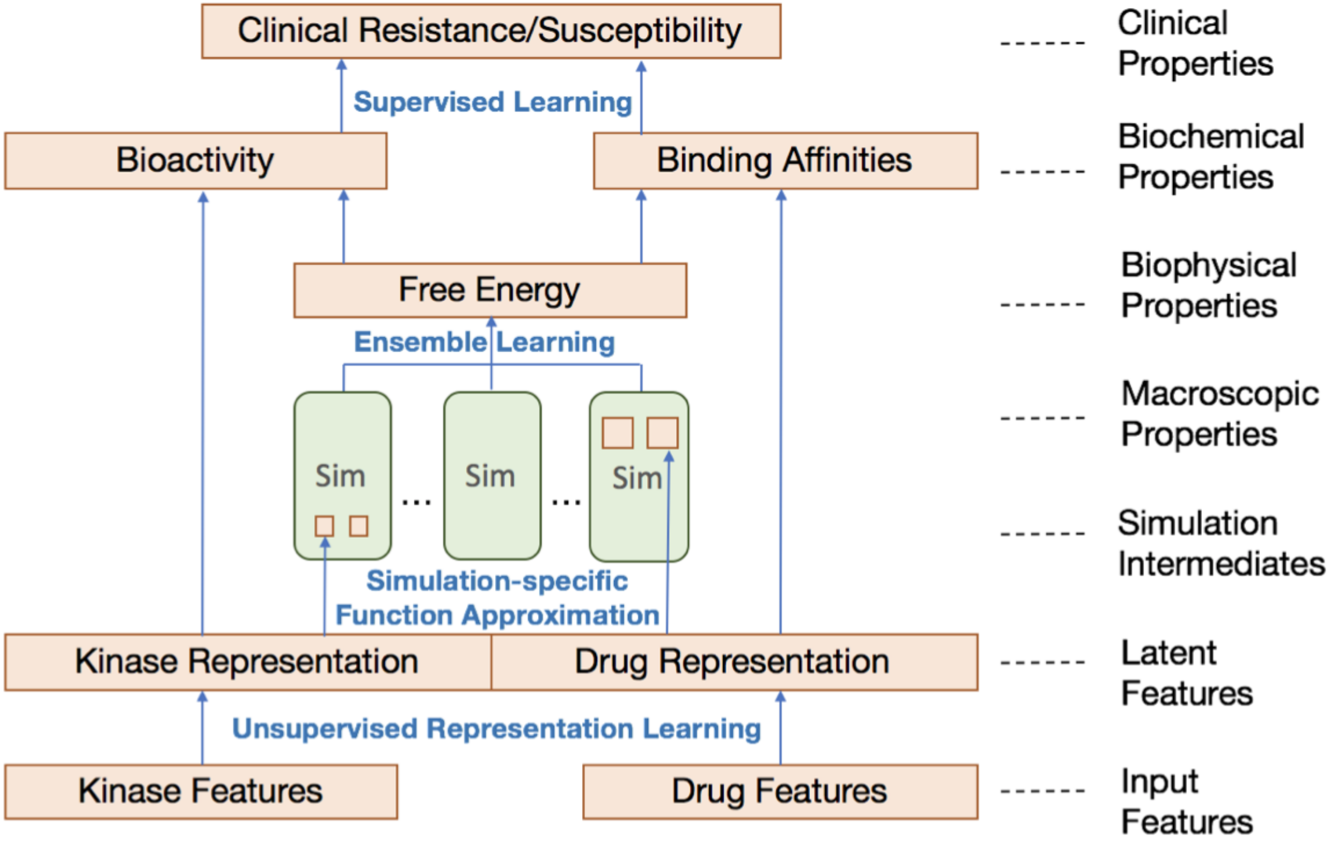
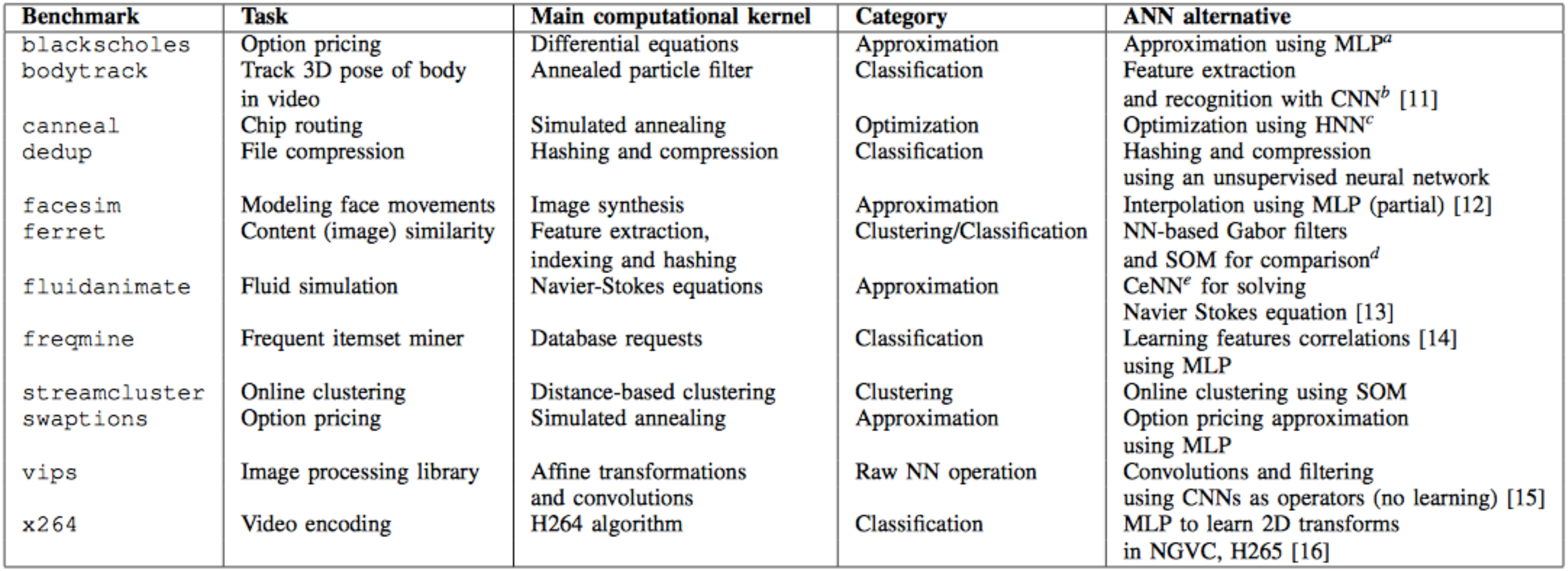


Figure 3: Overview of how data at all steps will be integrated using machine learning. The orange square boxes represent the three types of data in this project: kinases, drugs, and their interactions at various levels. The green rounded boxes denote the variety of MD simulations for free energy calculation. Each blue arrow represents an ML model; they combine in a joint predictive model that integrates all datasets.

The Critical Connections II

- Embedding Machine Learning into Simulations
  - Replacing explicit first principles models with learned functions
  - Faster, Lower Power, Lower Accuracy(?)
  - Functions in simulations accessing ML models at high throughput
  - On node invocation of dozens or hundreds of models millions of times per second?
  - Ex. Nowcasting in Weather

Algorithm Approximation



Benchmark	Task	Main computational kernel	Category	ANN alternative
blackscholes	Option pricing	Differential equations	Approximation	Approximation using MLP[a]
bodytrack	Track 3D pose of body in video	Annealed particle filter	Classification	Feature extraction and recognition with CNN[b] [11]
canneal	Chip routing	Simulated annealing	Optimization	Optimization using HNN[c]
dedup	File compression	Hashing and compression	Classification	Hashing and compression using an unsupervised neural network
facesim	Modeling face movements	Image synthesis	Approximation	Interpolation using MLP (partial) [12]
ferret	Content (image) similarity	Feature extraction, indexing and hashing	Clustering/Classification	NN-based Gabor filters and SOM for comparison[d]
fluidanimate	Fluid simulation	Navier-Stokes equations	Approximation	CeNN[e] for solving Navier Stokes equation [13]
freqmine	Frequent itemset miner	Database requests	Classification	Learning features correlations [14] using MLP
streamcluster	Online clustering	Distance-based clustering	Clustering	Online clustering using SOM
swaptions	Option pricing	Simulated annealing	Approximation	Option pricing approximation using MLP
vips	Image processing library	Affine transformations and convolutions	Raw NN operation	Convolutions and filtering using CNNs as operators (no learning) [15]
x264	Video encoding	H264 algorithm	Classification	MLP to learn 2D transforms in NGVC, H265 [16]

Neural Acceleration for General-Purpose Approximate Programs

Hadi Esmaeilzadeh Adrian Sampson Luis Ceze Doug Burger∗

University of Washington ∗Microsoft Research

Replacing Imperative Code with NN Computed Approximations

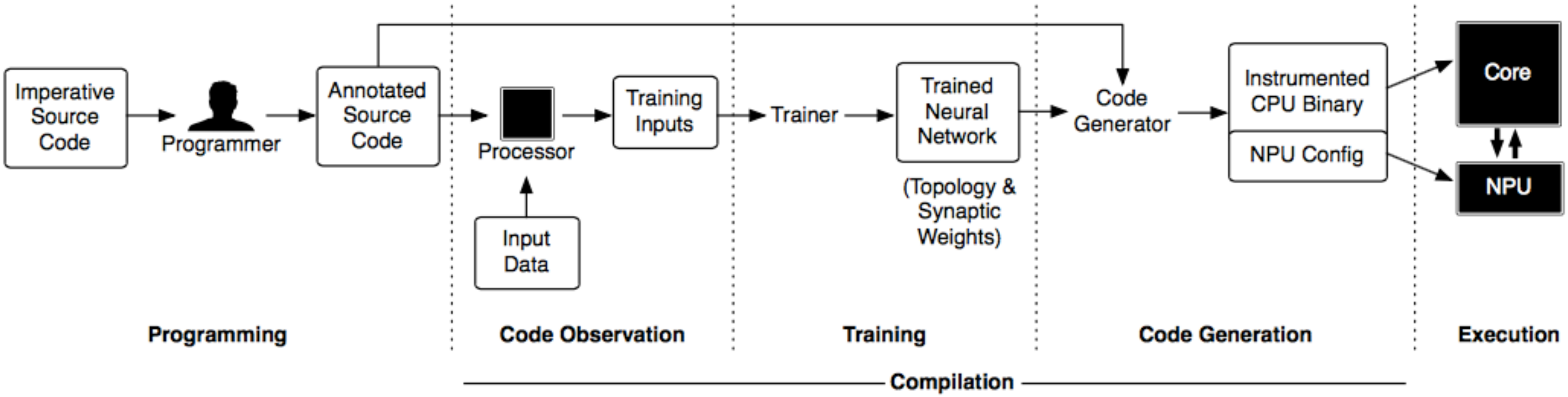


Figure 1: The Parrot transformation at a glance: from annotated code to accelerated execution on an NPU-augmented core.

Neural Acceleration for General-Purpose Approximate Programs

Hadi Esmaeilzadeh Adrian Sampson Luis Ceze Doug Burger*
University of Washington *Microsoft Research

2.3x Speedup, 3x Power Reduction, ~7% Error

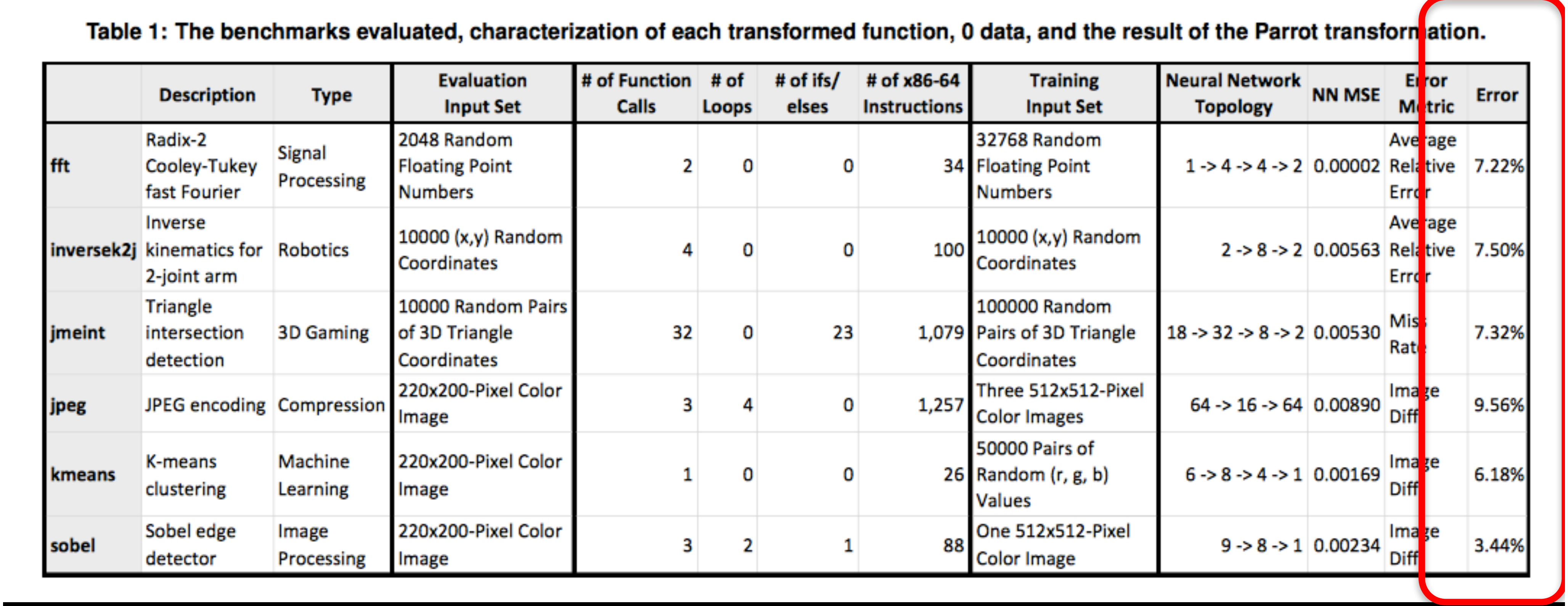


Table 1: The benchmarks evaluated, characterization of each transformed function, 0 data, and the result of the Parrot transformation.

	Description	Type	Evaluation Input Set	# of Function Calls	# of Loops	# of ifs/ elses	# of x86-64 Instructions	Training Input Set	Neural Network Topology	NN MSE	Error Metric	Error
fft	Radix-2 Cooley-Tukey fast Fourier	Signal Processing	2048 Random Floating Point Numbers	2	0	0	34	32768 Random Floating Point Numbers	1 -> 4 -> 4 -> 2	0.00002	Average Relative Error	7.22%
inversek2j	Inverse kinematics for 2-joint arm	Robotics	10000 (x,y) Random Coordinates	4	0	0	100	10000 (x,y) Random Coordinates	2 -> 8 -> 2	0.00563	Average Relative Error	7.50%
jmeint	Triangle intersection detection	3D Gaming	10000 Random Pairs of 3D Triangle Coordinates	32	0	23	1,079	100000 Random Pairs of 3D Triangle Coordinates	18 -> 32 -> 8 -> 2	0.00530	Miss Rate	7.32%
jpeg	JPEG encoding	Compression	220x200-Pixel Color Image	3	4	0	1,257	Three 512x512-Pixel Color Images	64 -> 16 -> 64	0.00890	Image Diff	9.56%
kmeans	K-means clustering	Machine Learning	220x200-Pixel Color Image	1	0	0	26	50000 Pairs of Random (r, g, b) Values	6 -> 8 -> 4 -> 1	0.00169	Image Diff	6.18%
sobel	Sobel edge detector	Image Processing	220x200-Pixel Color Image	3	2	1	88	One 512x512-Pixel Color Image	9 -> 8 -> 1	0.00234	Image Diff	3.44%

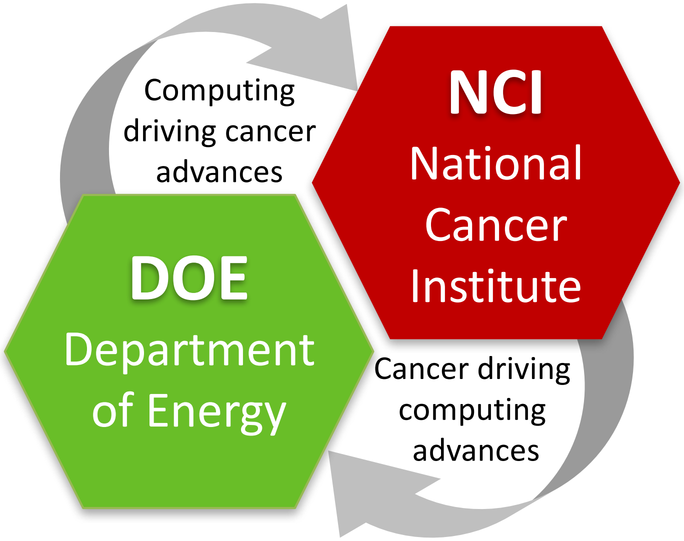
Neural Acceleration for General-Purpose Approximate Programs

Hadi Esmaeilzadeh Adrian Sampson Luis Ceze Doug Burger*
University of Washington *Microsoft Research

Launched in November 2015

Joint Design of Advanced Computing Solutions for Cancer

DOE-NCI partnership to advance cancer research and high performance computing in the U.S.

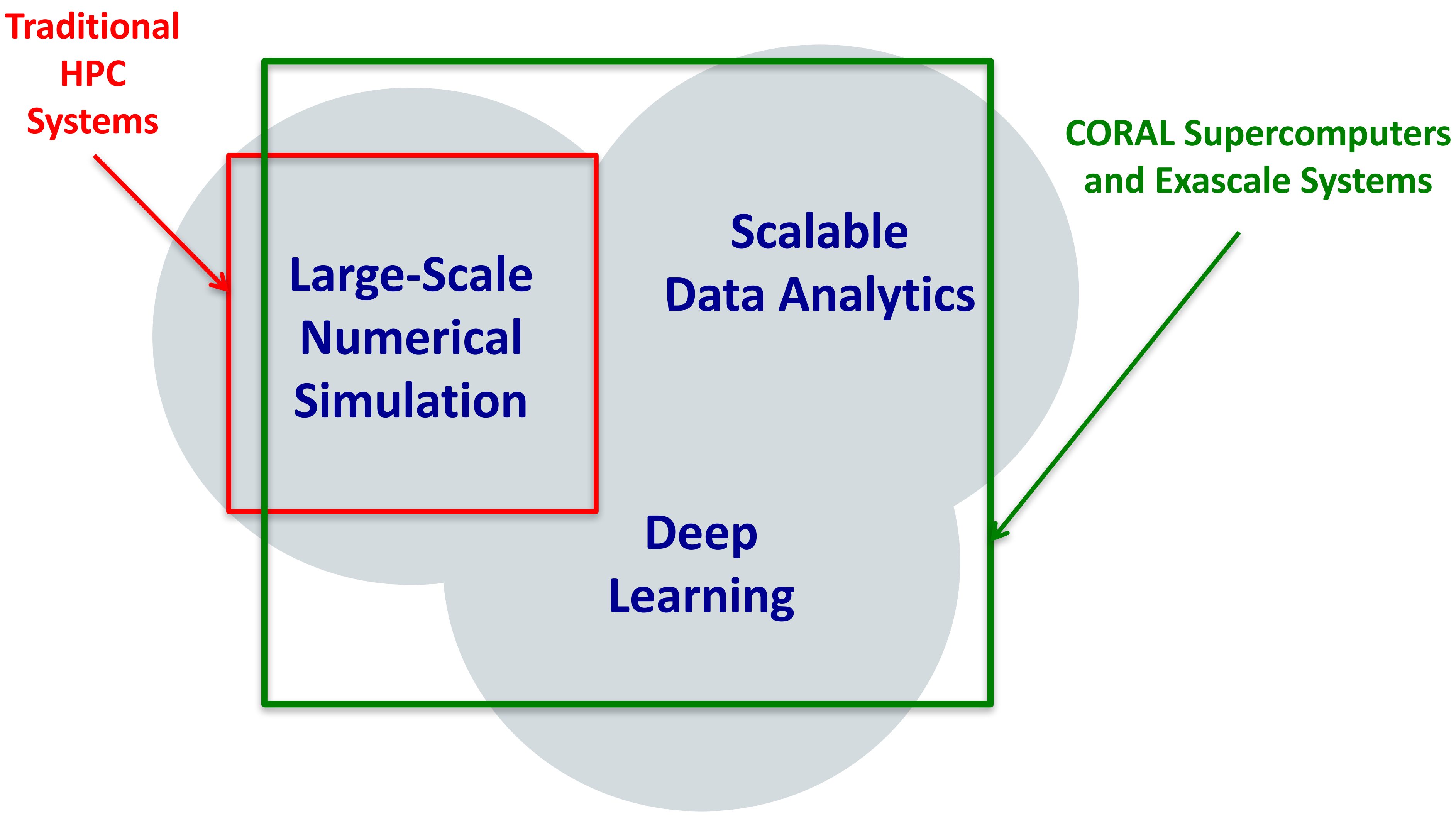

Computing driving cancer advances
NCI
National Cancer Institute
DOE
Department of Energy
Cancer driving computing advances

DOE Secretary of Energy

Director of the National Cancer Institute

U.S. DEPARTMENT OF ENERGY | NIH NATIONAL CANCER INSTITUTE

DOE Objective: Dirve Integration of Simulation, Data Analytics and Machine Learning



Traditional HPC Systems

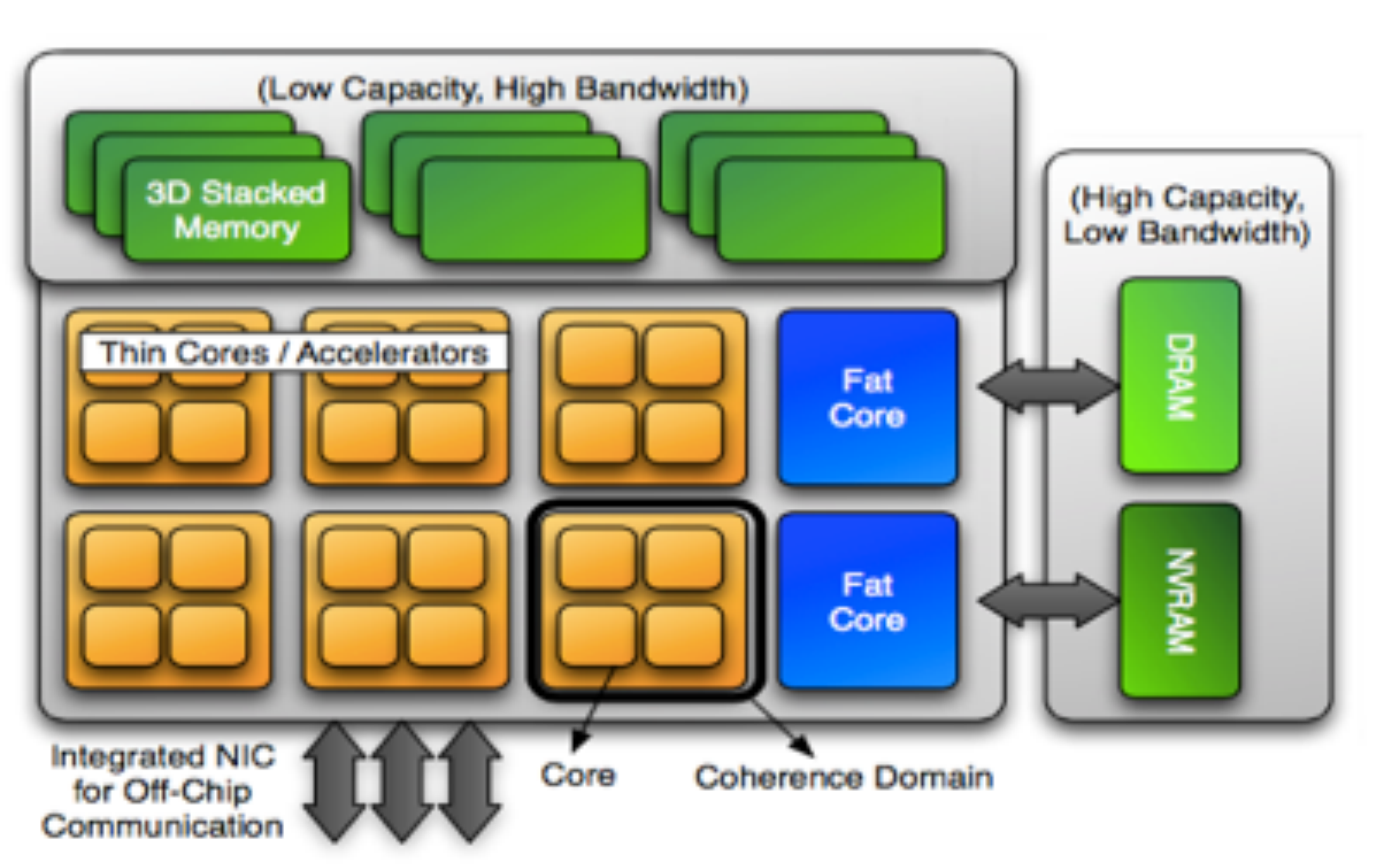
Large-Scale Numerical Simulation

Scalable Data Analytics

Deep Learning

CORAL Supercomputers and Exascale Systems

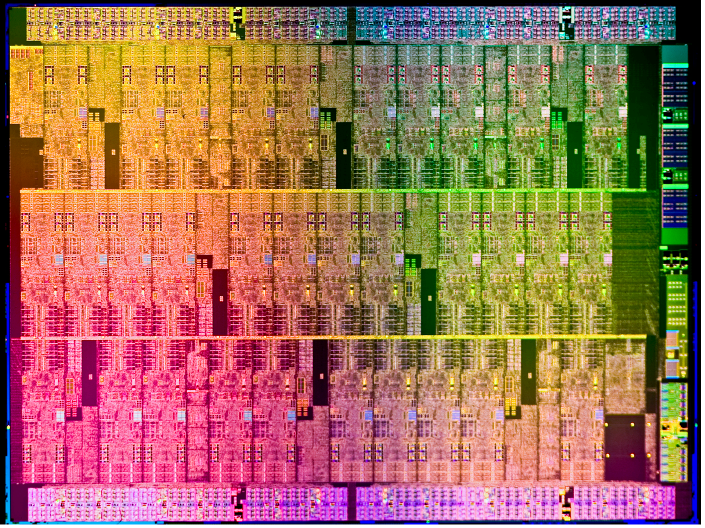
Exascale Node Concept Space


(Low Capacity, High Bandwidth)
3D Stacked Memory
(High Capacity, Low Bandwidth)
Thin Cores / Accelerators
Fat Core
Fat Core
DRAM
NVRAM
Integrated NIC for Off-Chip Communication
Core
Coherence Domain

Abstract Machine Models and Proxy Architectures for Exascale Computing Rev 1.1
Sandia National Laboratory and Lawrence Berkeley National Laboratory

Leverage Resources on the Die, in Package or on the Node

- Local high-bandwidth memory stacks
- Node based non-volitile memory
- High-Bandwidth Low Latency Fabric
- General Purpose Cores
- Dynamic Power Management



What Kind of Accelerator(s) to Add?

- Vector Processors
- Data Flow Engines
- Patches of FPGA
- Many “Nano” Cores (< 5 M Tr each?)
- Neuromorphic Cores
- CNN Cores
- Tensor Engines
- Other Machine Learning Cores?

Hardware and systems architectures are emerging for supporting deep learning

- CPUs
  - AVX, VNNI, KNL, KNM, KNH, …
- GPUs
  - Nvidia P100, V100, AMD Instinct, Baidu GPU, …
- ASICs
  - Nervana, DianNao, Eyeriss, GraphCore, TPU, DLU, …
- FPGA
  - Arria 10, Stratix 10, Falcon Mesa, …
- Neuromorphic
  - True North, Zeroth, N1, …

Aurora 21

- Argonne’s Exascale System
- Balanced architecture to support three pillars
  - Large-scale Simulation (PDEs, traditional HPC)
  - Data Intensive Applications (science pipelines)
  - Deep Learning and Emerging Science AI
- Enable integration and embedding of pillars
- Integrated computing, acceleration, storage
- Towards a common software stack

Argonne Targets for Exascale

Simulation Applications

- Materials Science
- Cosmology
- Molecular Dynamics
- Nuclear Reactor Modeling
- Combustion
- Quantum Computer Simulation
- Climate Modeling
- Power Grid
- Discrete Event Simulation
- Fusion Reactor Simulation
- Brain Simulation
- Transportation Networks

Big Data Applications

- APS Data Analysis
- HEP Data Analysis
- LSST Data Analysis
- SKA Data Analysis
- Metagenome Analysis
- Battery Design Search
- Graph Analysis
- Virtual Compound Library
- Neuroscience Data Analysis
- Genome Pipelines

Deep Learning Applications

- Drug Response Prediction
- Scientific Image Classification
- Scientific Text Understanding
- Materials Property Design
- Gravitational Lens Detection
- Feature Detection in 3D
- Street Scene Analysis
- Organism Design
- State Space Prediction
- Persistent Learning
- Hyperspectral Patterns

Differing Requirements?

Simulation Applications

- **64bit floating point**
- **Memory Bandwith**
- **Random Access to Memory**
- **Sparse Matrices**
- **Distributed Memory jobs**
- **Synchronous I/O multinode**
- **Scalability Limited Comm**
- **Low Latency High Bandwidth**
- **Large Coherency Domains help sometimes**
- **O typically greater than I**
- **O rarely read**
- **Output is data**

Big Data Applications

- 64 bit and Integer important
- Data analysis Pipelines
- DB including No SQL
- MapReduce/SPARK
- Millions of jobs
- I/O bandwidth limited
- Data management limited
- Many task parallelism
- Large-data in and Large-data out
- I and O both important
- O is read and used
- Output is data

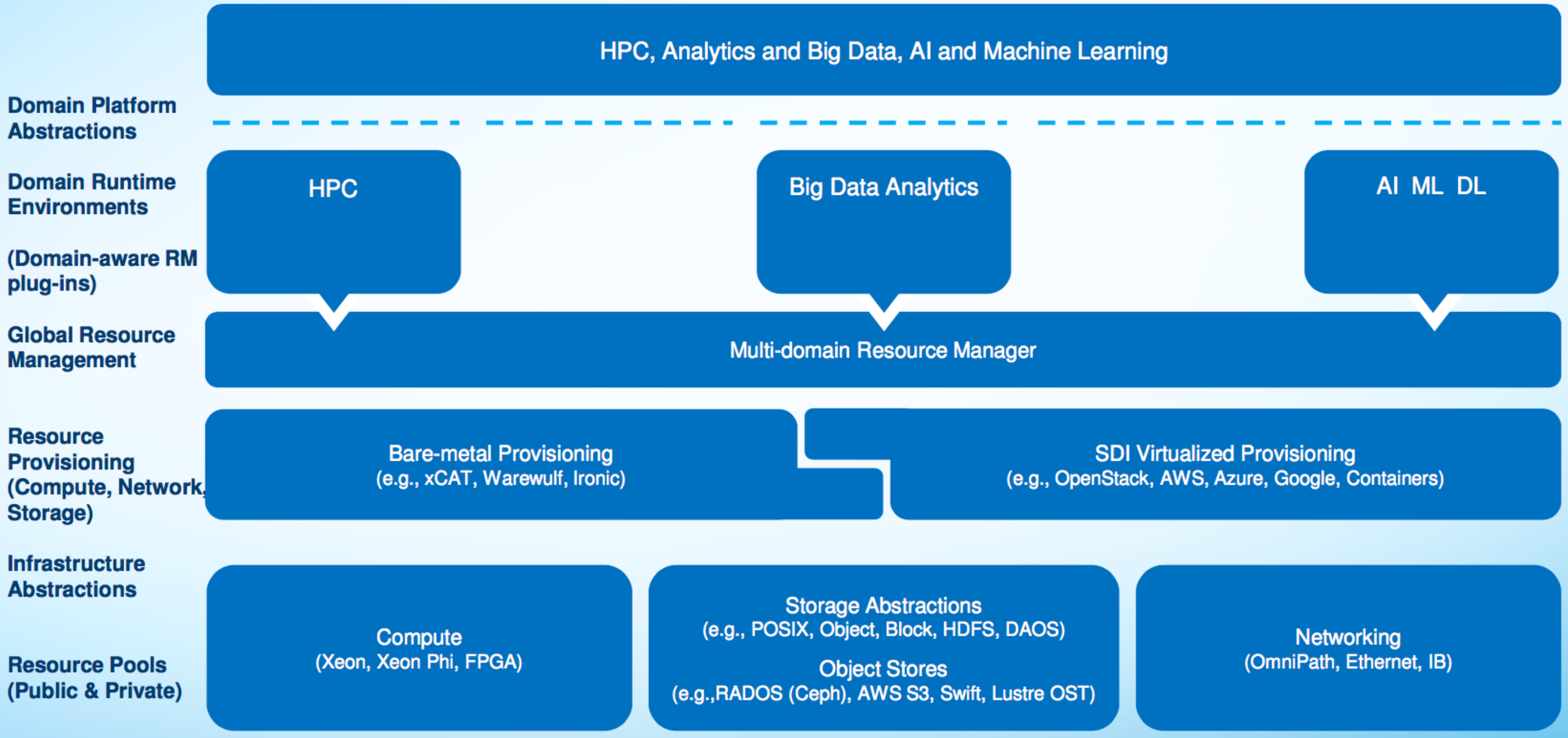
Deep Learning Applications

- Lower Precision (fp32, fp16)
- FMAC @ 32 and 16 okay
- Inferencing can be 8 bit (TPU)
- Scaled integer possible
- Training dominates dev
- Inference dominates pro
- Reuse of training data
- Data pipelines needed
- Dense FP typical SGEMM
- Small DFT, CNN
- Ensembles and Search
- Single Models Small
- I more important than O
- Output is models

Aurora 21 Exascale Software

- Single Unified stack with resource allocation and scheduling across all pillars and ability for frameworks and libraries to seamlessly compose
- Minimize data movement: keep permanent data in the machine via distributed persistent memory while maintaining availability requirements
- Support standard file I/O and path to memory coupled model for Sim, Data and Learning
- Isolation and reliability for multi-tenancy and combining workflows

Towards an Integrated Stack



Domain Platform Abstractions

HPC, Analytics and Big Data, AI and Machine Learning

Domain Runtime Environments

(Domain-aware RM plug-ins)

HPC

Big Data Analytics

AI ML DL

Global Resource Management

Multi-domain Resource Manager

Resource Provisioning (Compute, Network, Storage)

Bare-metal Provisioning
(e.g., xCAT, Warewulf, Ironic)

SDI Virtualized Provisioning
(e.g., OpenStack, AWS, Azure, Google, Containers)

Infrastructure Abstractions

Resource Pools (Public & Private)

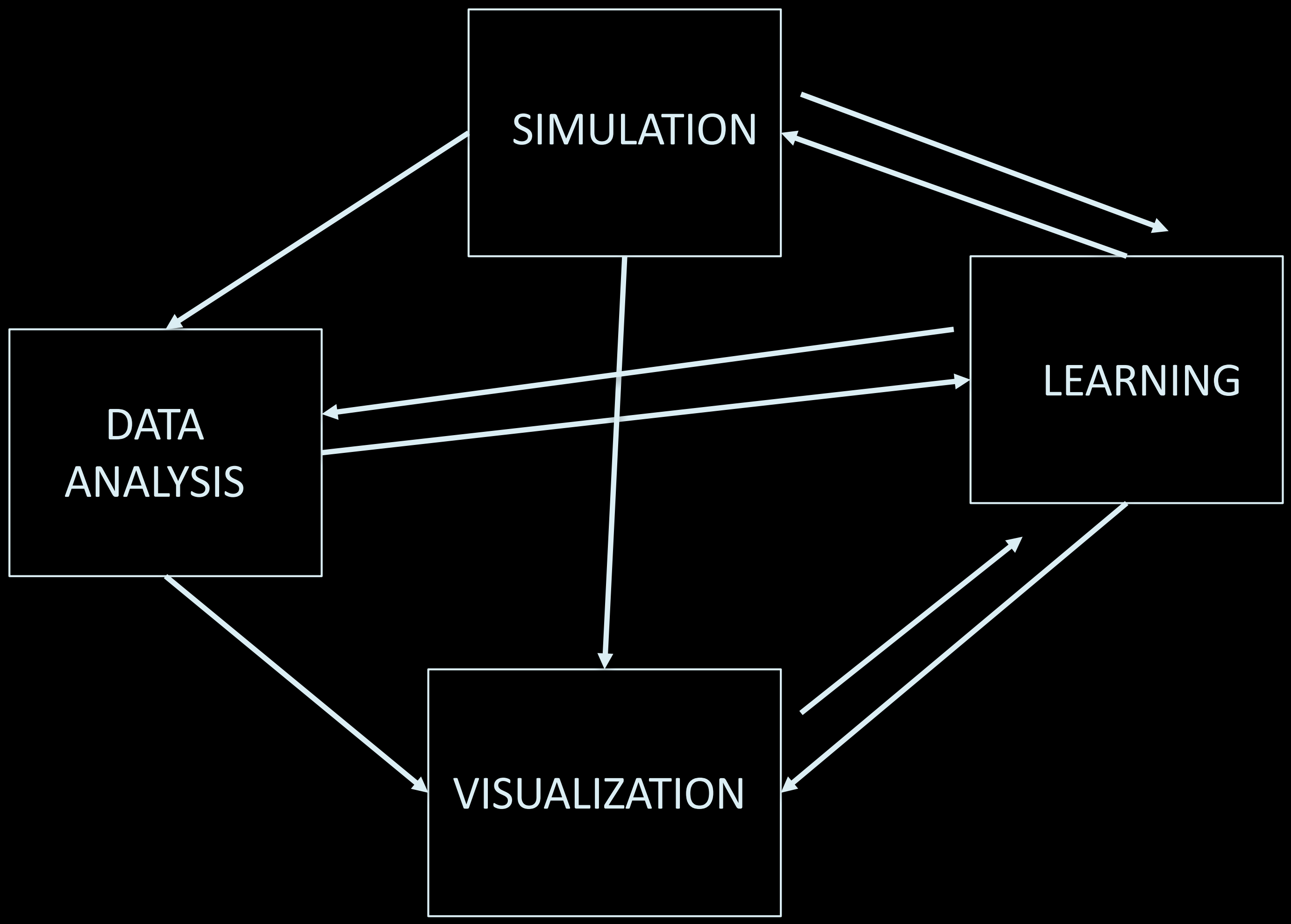
Compute
(Xeon, Xeon Phi, FPGA)

Storage Abstractions
(e.g., POSIX, Object, Block, HDFS, DAOS)

Object Stores
(e.g.,RADOS (Ceph), AWS S3, Swift, Lustre OST)

Networking
(OmniPath, Ethernet, IB)

The New HPC "Paradigm"



Acknowledgements

Many thanks to DOE, NSF, NIH, DOD, ANL, UC, Moore Foundation, Sloan Foundation, Apple, Microsoft, Cray, Intel, NVIDIA and IBM for supporting our research group over the years

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

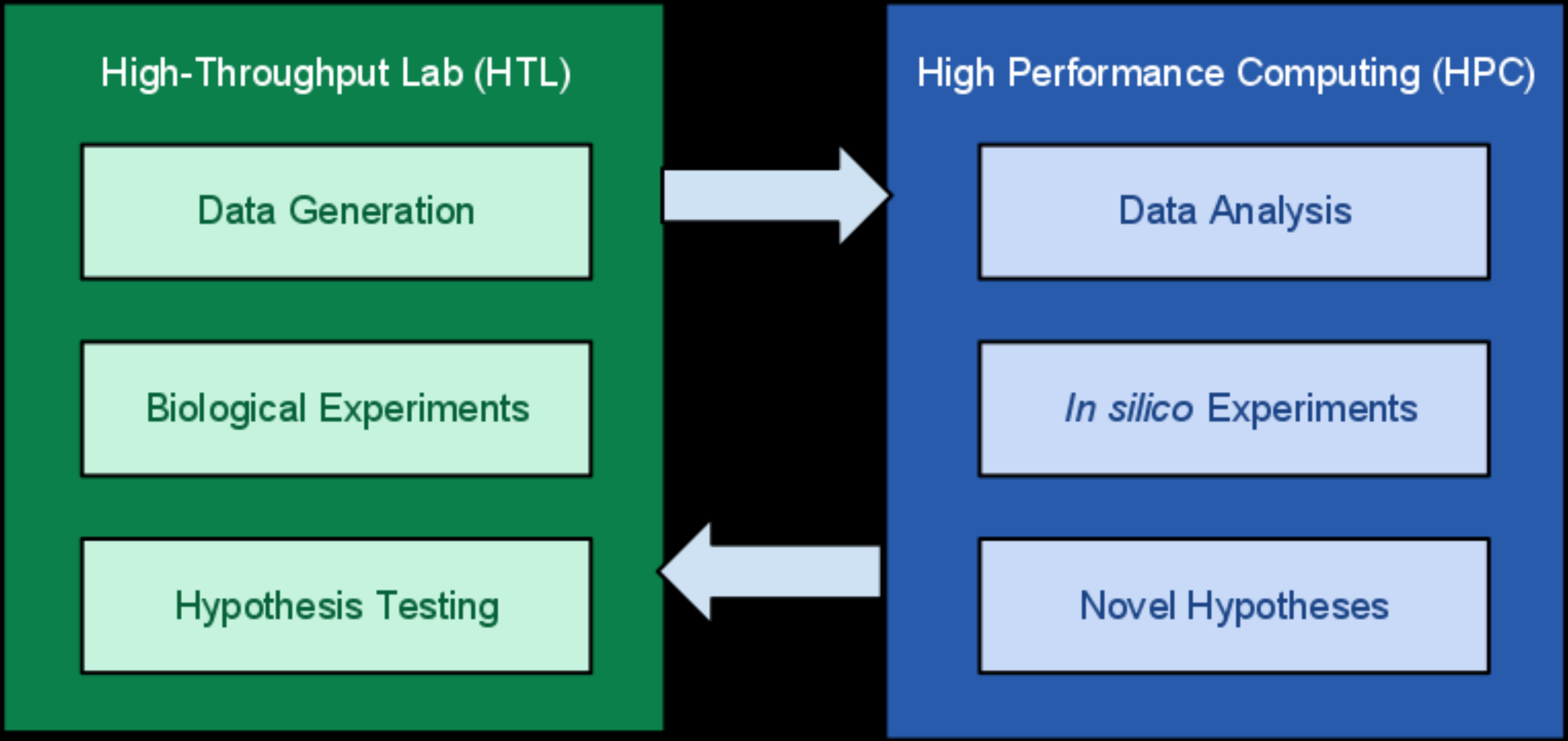
NSF

NATIONAL INSTITUTES OF HEALTH

NVIDIA. IBM CRAY intel

END!

Our Vision
Automate and Accelerate



High-Throughput Lab (HTL)

Data Generation

Biological Experiments

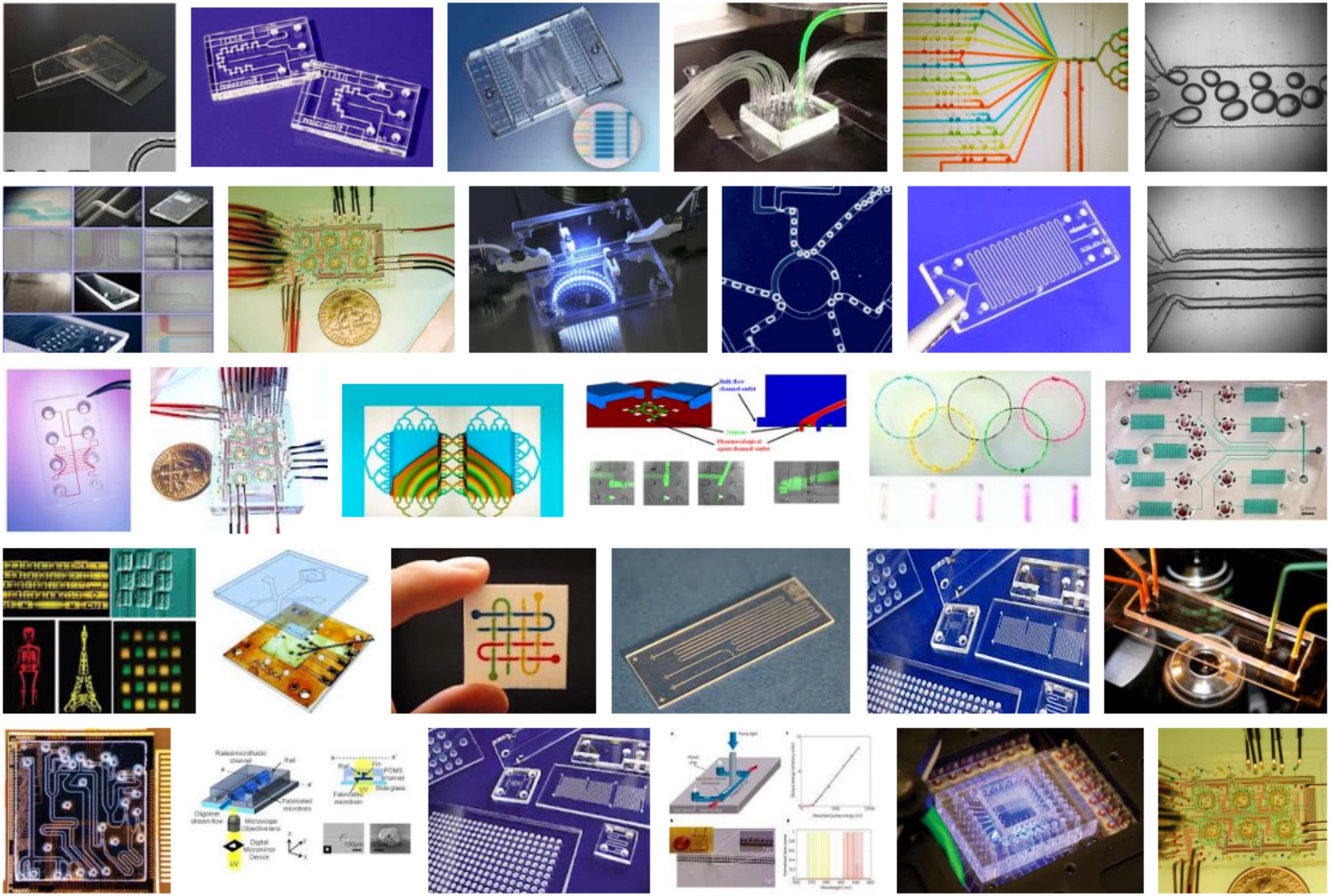
Hypothesis Testing

High Performance Computing (HPC)

Data Analysis

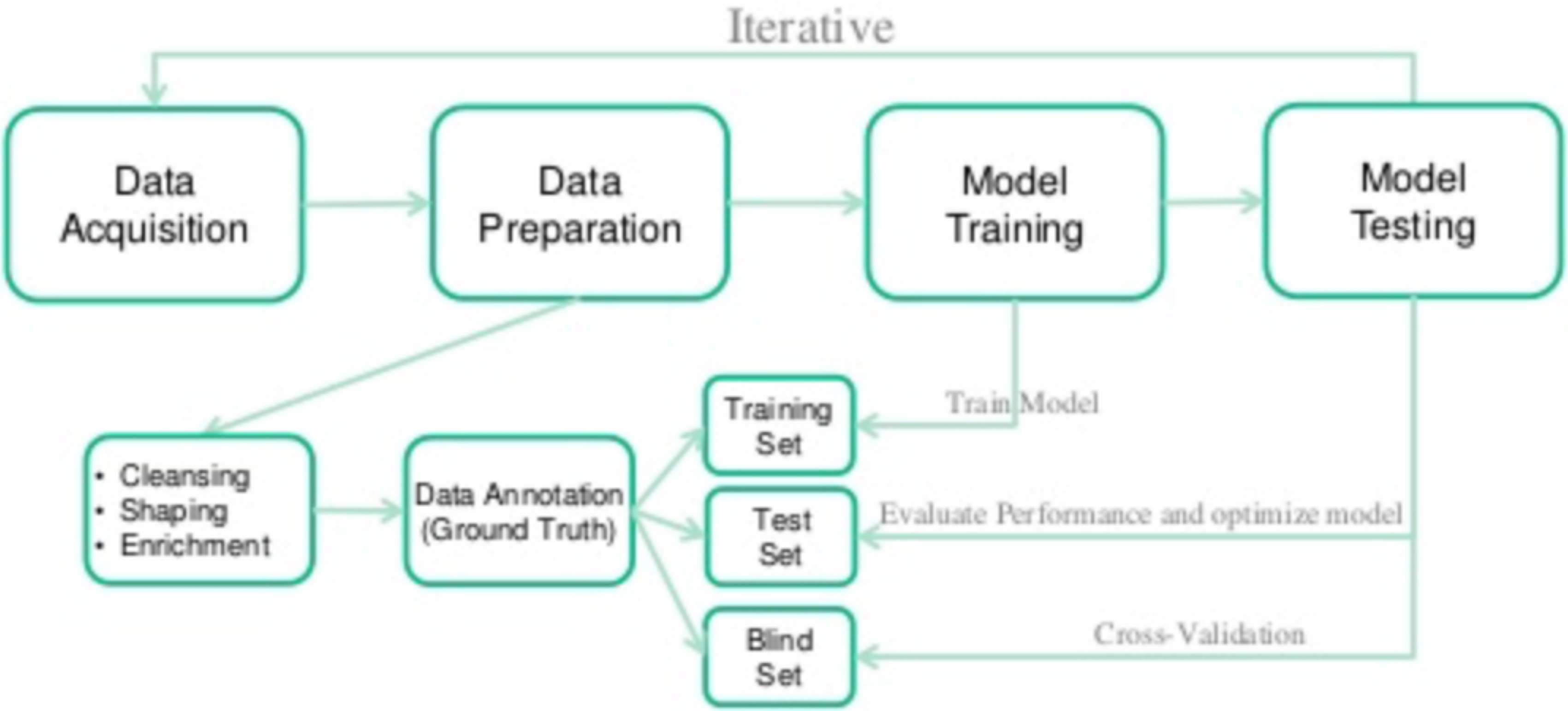
In silico Experiments

Novel Hypotheses

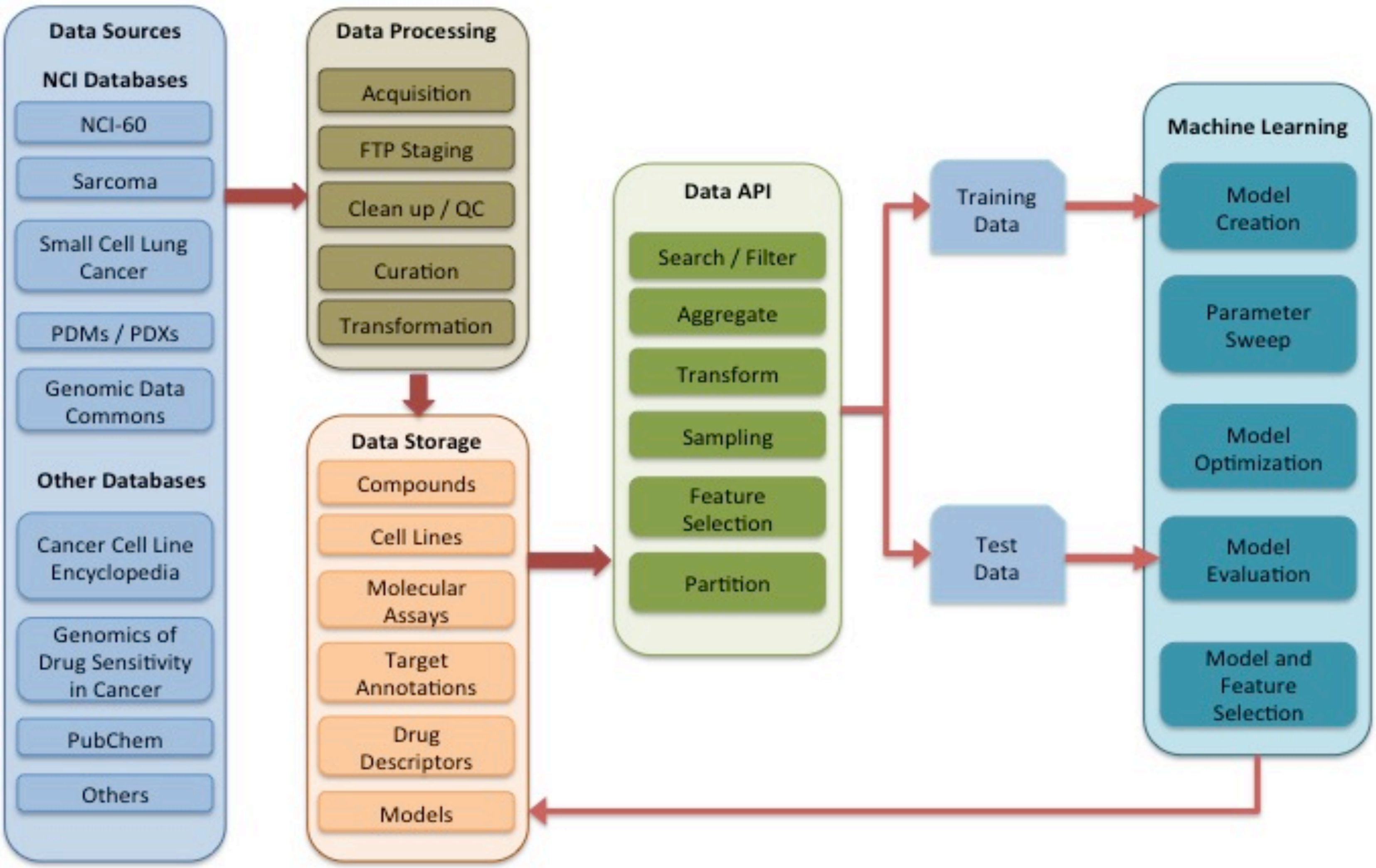

micronit
micronit
Bulk flow channel outlet
Pharmacological agent channel outlet
5mm
Railed microfluidic channel
Rail
Fabricated microtrain
Oligomer stream flow
Microscope Objective lens
Digital Micromirror Device
UV
Fin
PDMS channel
Slide glass
Fabricated microtrain
PDMS

The CANDLE Exascale Project

Typical Machine Learning Flow diagram



Drug Response **CANDLE General Workflow**



Data Sources

NCI Databases

- NCI-60
- Sarcoma
- Small Cell Lung Cancer
- PDMs / PDXs
- Genomic Data Commons

Other Databases

- Cancer Cell Line Encyclopedia
- Genomics of Drug Sensitivity in Cancer
- PubChem
- Others

Data Processing

- Acquisition
- FTP Staging
- Clean up / QC
- Curation
- Transformation

Data Storage

- Compounds
- Cell Lines
- Molecular Assays
- Target Annotations
- Drug Descriptors
- Models

Data API

- Search / Filter
- Aggregate
- Transform
- Sampling
- Feature Selection
- Partition

Training Data

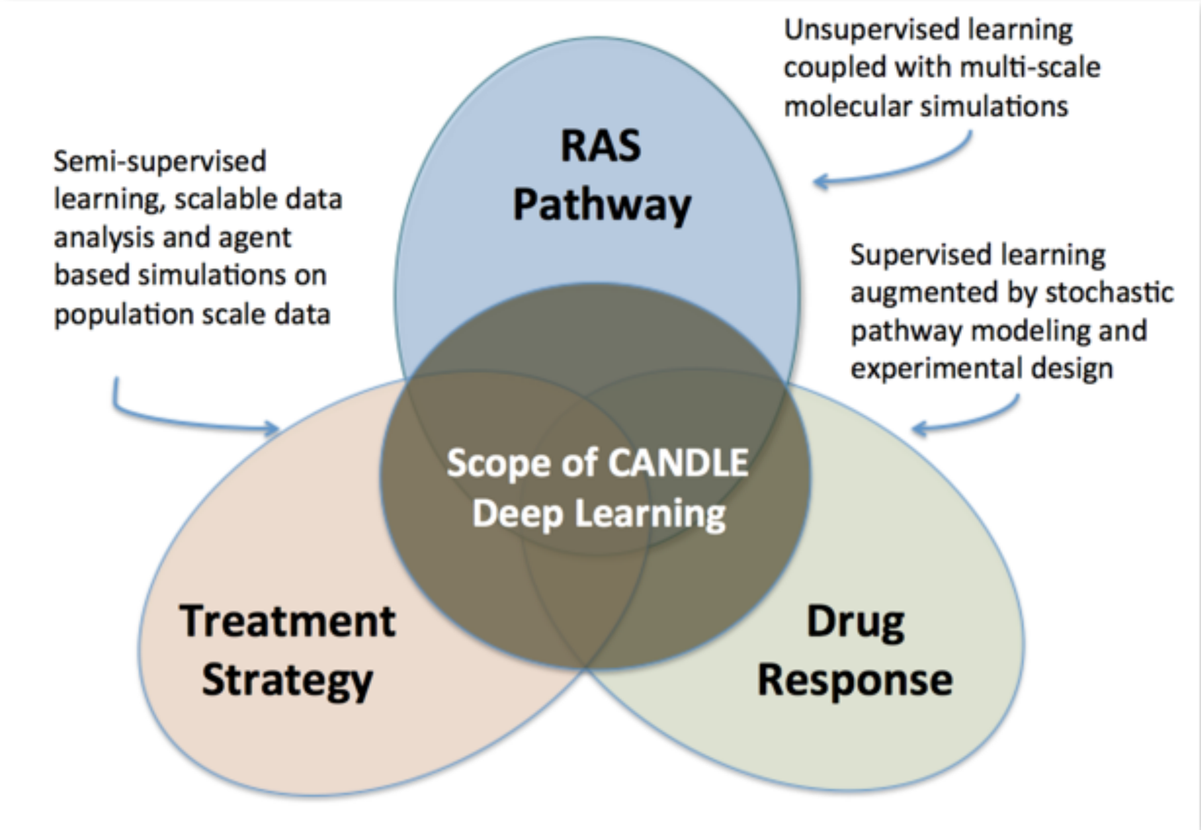
Test Data

Machine Learning

- Model Creation
- Parameter Sweep
- Model Optimization
- Model Evaluation
- Model and Feature Selection

Cancer Data Processing, Storage and Machine Learning Workflow

ECP-CANDLE : CANcer Distributed Learning Environment


Unsupervised learning coupled with multi-scale molecular simulations
RAS Pathway
Semi-supervised learning, scalable data analysis and agent based simulations on population scale data
Supervised learning augmented by stochastic pathway modeling and experimental design
Scope of CANDLE Deep Learning
Treatment Strategy
Drug Response

CANDLE Goals

Develop an exascale deep learning environment for cancer

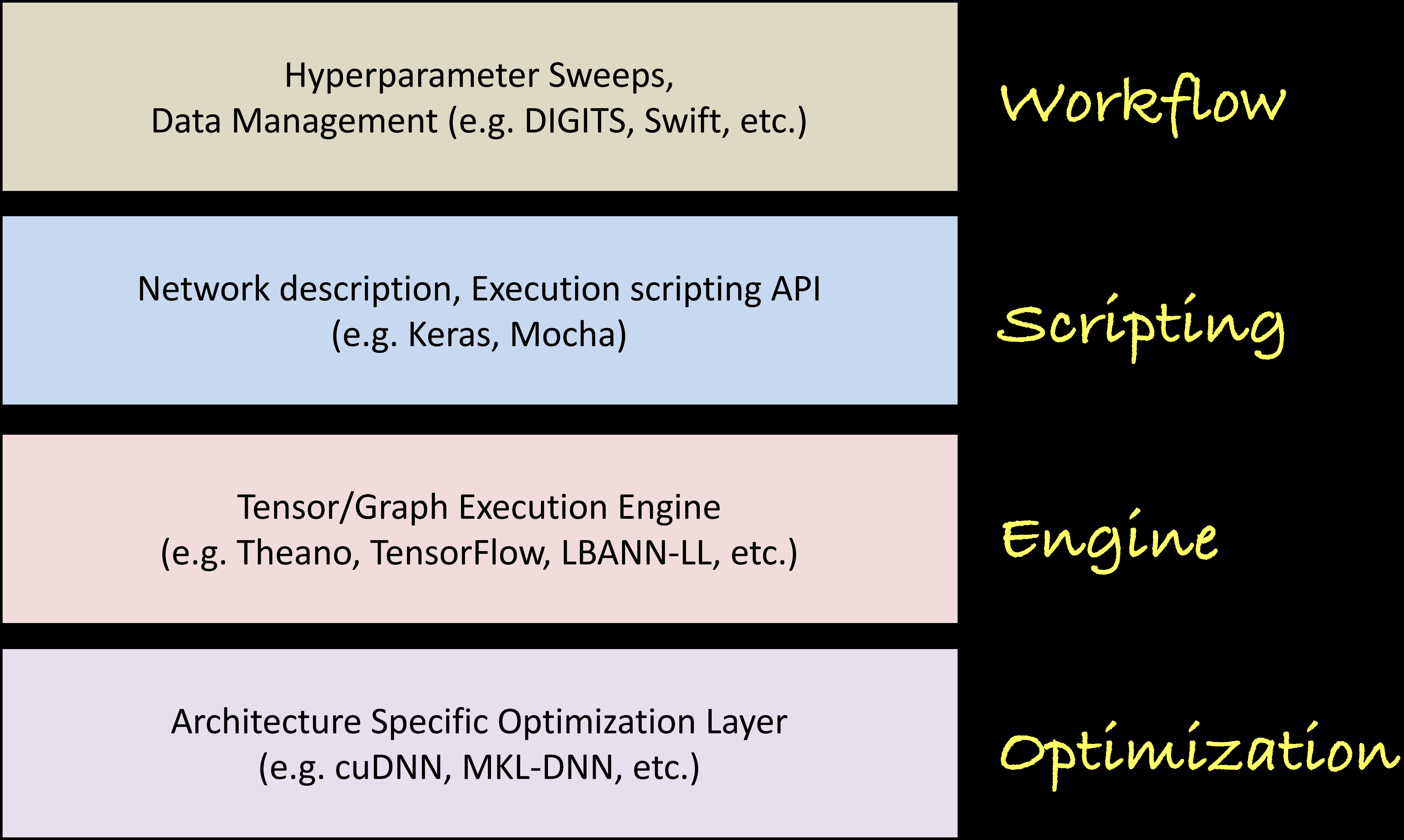
Building on open source Deep learning frameworks

Optimization for CORAL and exascale platforms

Support all three pilot project needs for deep

Collaborate with DOE computing centers, HPC vendors and ECP co-design and software technology projects

CANDLE Software Stack



Hyperparameter Sweeps,
Data Management (e.g. DIGITS, Swift, etc.)

Workflow

Network description, Execution scripting API
(e.g. Keras, Mocha)

Scripting

Tensor/Graph Execution Engine
(e.g. Theano, TensorFlow, LBANN-LL, etc.)

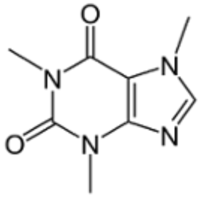
Engine

Architecture Specific Optimization Layer
(e.g. cuDNN, MKL-DNN, etc.)

Optimization

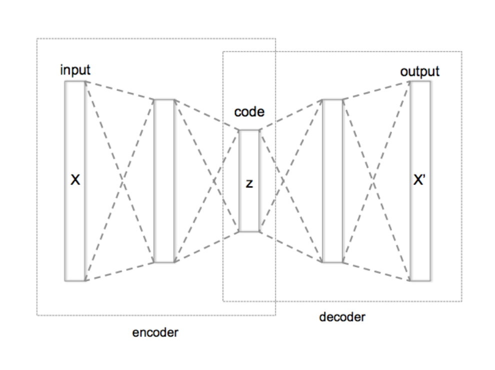
DL Frameworks “Tensor Engines”

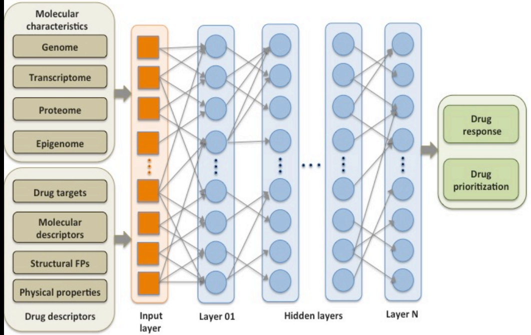
- **TensorFlow** (c++, symbolic diff+)
- **Theano** (c++, symbolic diff+)
- **Neon** (integrated) (python + GPU, symbolic diff+)
- **Mxnet** (integrated) (c++)
- **LBANN** (c++, aimed at scalable hardware)
- **pyTorch7 TH Tensor** (c layer, symbolic diff-, pgks)
- **Caffe** (integrated) (c++, symbolic diff-)
- **Mocha** backend (julia + GPU)
- **CNTK** backend (microsoft) (c++)
- **PaddlePaddle** (Baidu) (python, c++, GPU)

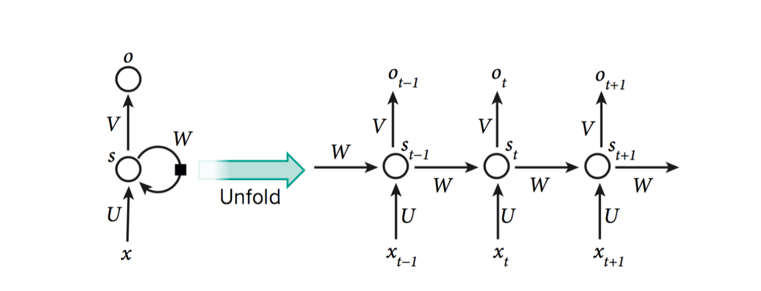


CANDLE Benchmarks.. Representative problems

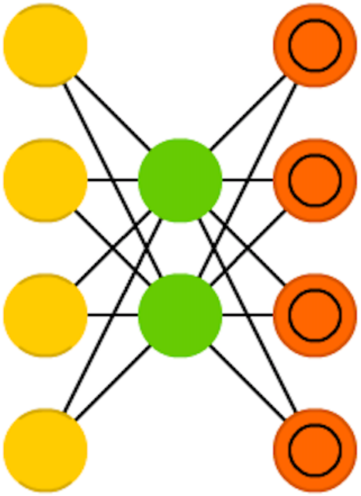
- Variational AutoEncoder
  - Learning (non-linear) features of core data types
- AutoEncoder
  - Molecular dynamics trajectory state detection
- MLP+LCNN Classification
  - Cancer type from gene expression/SNPs
- MLP+CNN Regression
  - Drug response (gene exp, descriptors)
- CNN
  - Cancer pathology report term extraction
- RNN-LSTM
  - Cancer pathology report text analysis
- RNN-LSTM
  - Molecular dynamics simulation control



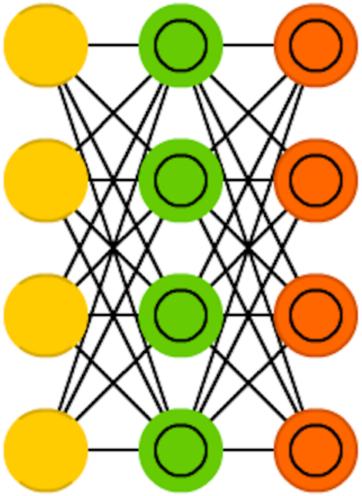




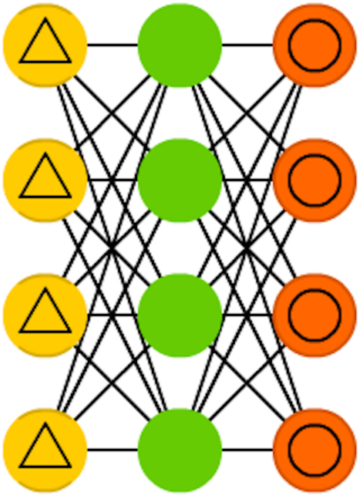
Auto Encoder (AE)



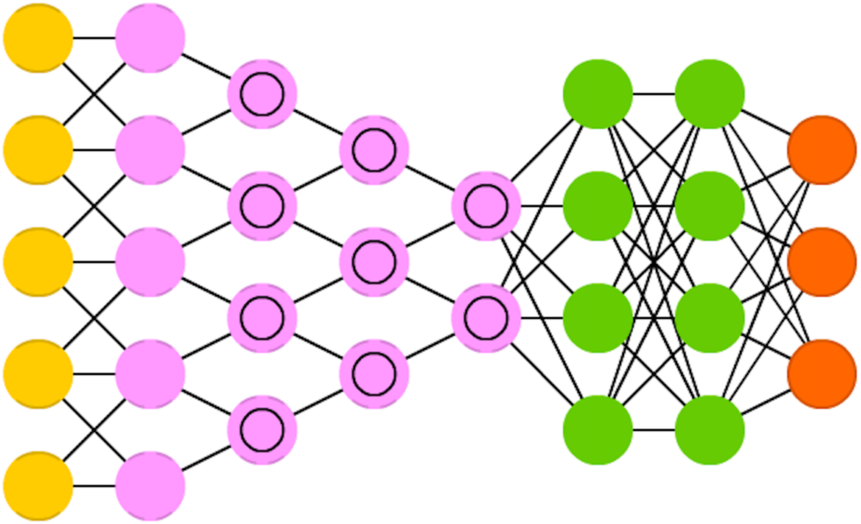
Variational AE (VAE)



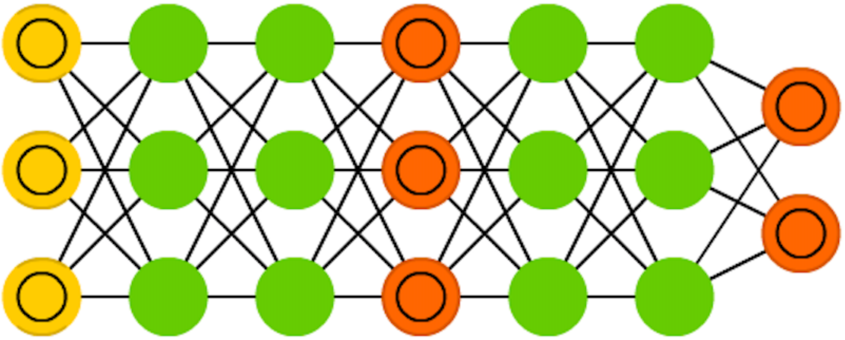
Denoising AE (DAE)



Deep Convolutional Network (DCN)



Generative Adversarial Network (GAN)



Progress in Deep Learning for Cancer

- **AutoEncoders** – learning data representations for classificaiton and prediction of drug response, molecular trajectories
- **VAEs and GANs** – generating data to support methods development, data augmentation and feature space algebra, drug candidate generation
- **CNNs** – type classification, drug response, outcomes prediction, drug resistance
- **RNNs** – sequence, text and molecular trajectories analysis
- **Multi-Task Learning** – terms (from text) and feature extraction (data), data translation (RNAseq <-> uArray)

CANDLE - FOM – Rate of Training

- **“Number of networks trained per day”**
  - – size and type of network, amount of training data, batch size, number of epochs, type of hardware
- **“Number of ‘weight’ updates/second”**
  - – Forward Pass + Backward Pass
- Training Rate = $\sum^{n}_{i=1} a_iR_i$ where R_i is the rate for our benchmark *i* and a_i is a weight

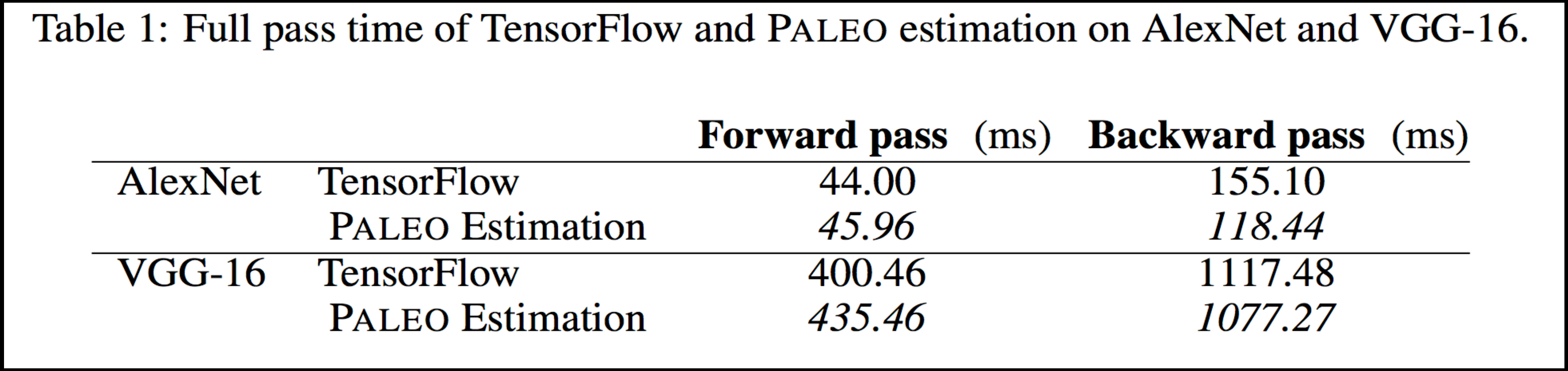


Table 1: Full pass time of TensorFlow and PALEO estimation on AlexNet and VGG-16.

		Forward pass (ms)	**Backward pass** (ms)
AlexNet	TensorFlow	44.00	155.10
	PALEO Estimation	*45.96*	*118.44*
VGG-16	TensorFlow	400.46	1117.48
	PALEO Estimation	*435.46*	*1077.27*

7 CANDLE Benchmarks

https://github.com/ECP-CANDLE

Benchmark Owners:

- P1: Fangfang Xia (ANL)
- P2: Brian Van Essen (LLNL)
- P3: Arvind Ramanathan (ORNL)

Benchmark	Type	Data	ID	OD	Sample Size	Size of Network	Additional (activation, layer types, etc.)
1. P1: B1 Autoencoder	MLP	RNA-Seq	10^5	10^5	15K	5 layers	Log2 (x+1) → [0,1] KPRM-UQ
2. P1: B2 Classifier	MLP	SNP → Type	10^6	4	15K	5 layers	Training Set Balance issues
3. P1: B3 Regression	MLP+LCN	expression; drug descs	10^5	1	3M	8 layers	Drug Response [-100, 100]
4. P2: B1 Autoencoder	MLP	MD K-RAS	10^5	10^2	10^6-10^8	5-8 layers	State Compression
5. P2: B2 RNN-LSTM	RNN-LSTM	MD K-RAS	10^5	3	10^6	4 layers	State to Action
6. P3: B1 RNN-LSTM	RNN-LSTM	Path reports	10^3	5	5K	1-2 layers	Dictionary 12K +30K
7. P3: B2 Classification	CNN	Path reports	10^4	10^2	10^5	5 layers	Biomarkers

Drug Response

RAS Pathways

Patient Trajectories

U.S. DEPARTMENT OF ENERGY | NIH NATIONAL CANCER INSTITUTE

Typical Performance Experience

CANDLE - Predicting drug response of tumor samples

- MLP/CNN on Keras
- 7 layers, 30M - 500M parameters
- 200 GB input size
- 1 hour/epoch on DGX-1; 200 epochs take 8 days (200 GPU hrs)
- Hyperparameter search ~ 200,000 GPU hrs or 8M CPU hrs

Protein function classification in genome annotation

- Deep residual convolution network on Keras
- 50 layers
- 1 GB input size
- 20 minutes/epoch on DGX-1; 200 epochs take 3 days (72 GPU hrs)
- Hyperparameter search ~ 72,000 GPU hrs or 2.8M CPU hrs

Github and FTP

- ECP-CANDLE GitHub Organization:
- https://github.com/ECP-CANDLE

- ECP-CANDLE FTP Site:
- The FTP site hosts all the public datasets for the benchmarks from three pilots.
- http://ftp.mcs.anl.gov/pub/candle/public/

Things We Need

- Deep Learning Workflow Tools
- Data Management for Training Data and Models
- Performance Measurement, Modeling and Monitoring of Training Runs
- Deep Network Model Visualization
- Low-level Solvers, Optimization and Data Encoding
- Programming Models/Runtimes to support next generation Parallel Deep Learning with sparsity
- OS Support for High-Throughput Training