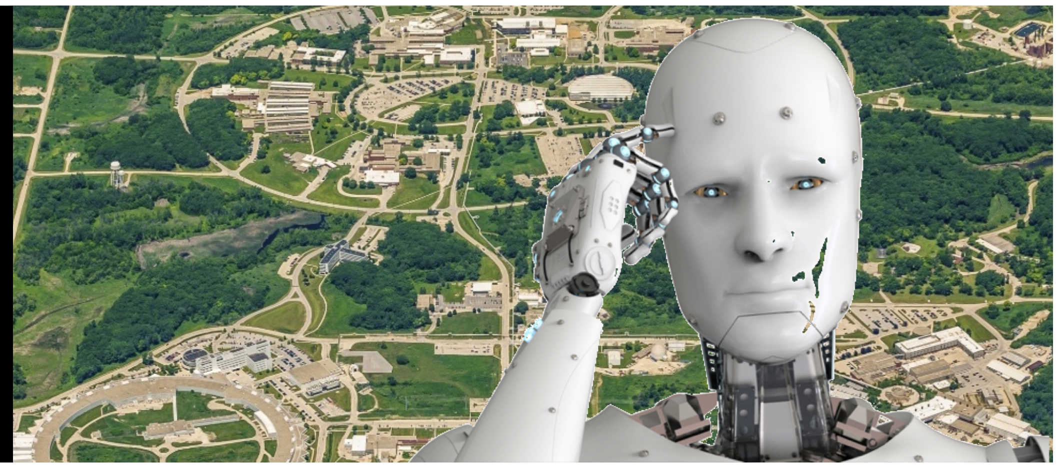
AI for Science @ Argonne



IAN FOSTER

Director, Data Science and Learning Division, Argonne National Laboratory
Professor of Computer Science, University of Chicago
foster@anl.gov

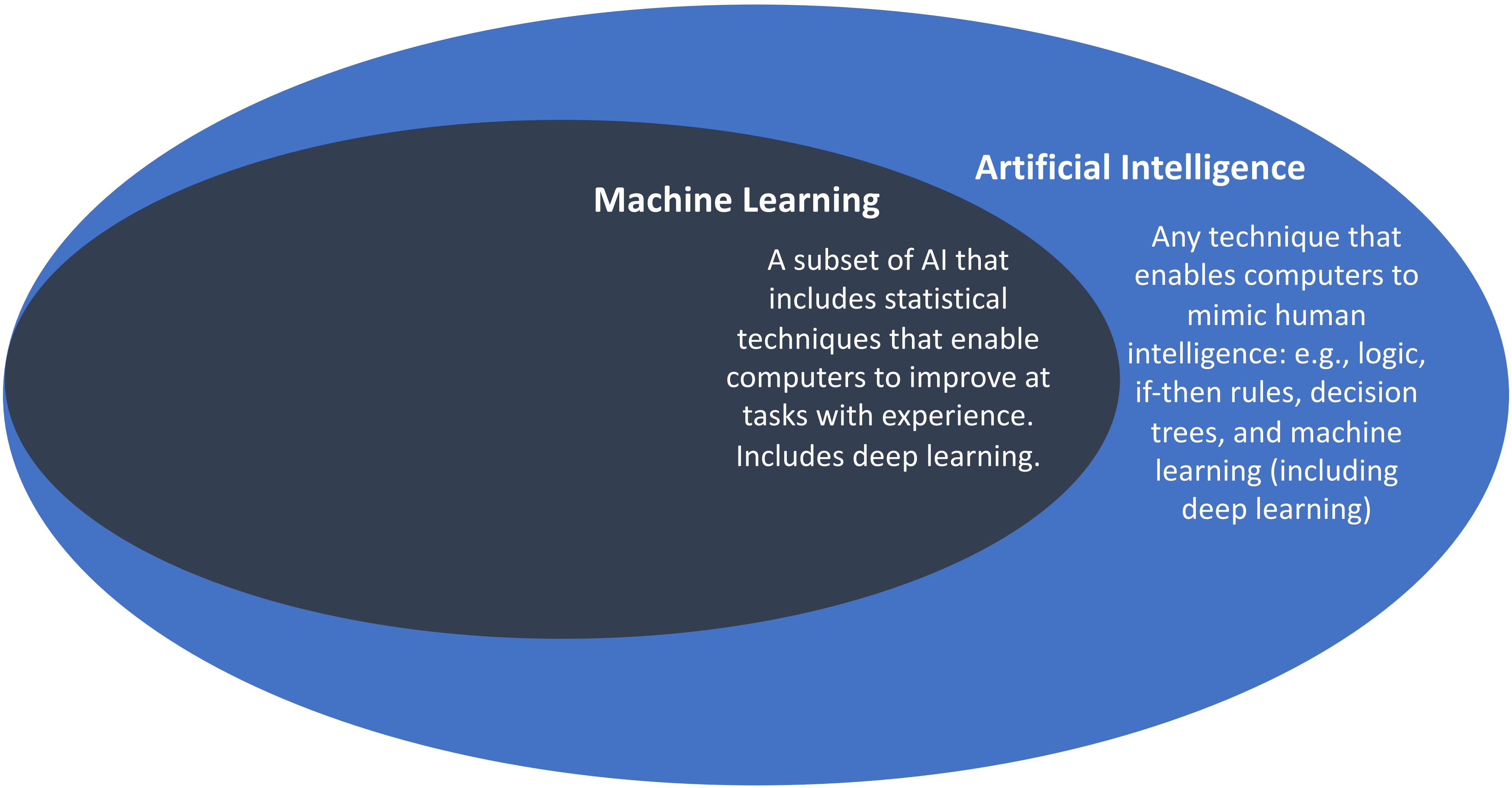
Fermilab, January 14, 2020

Briefly: What is AI?

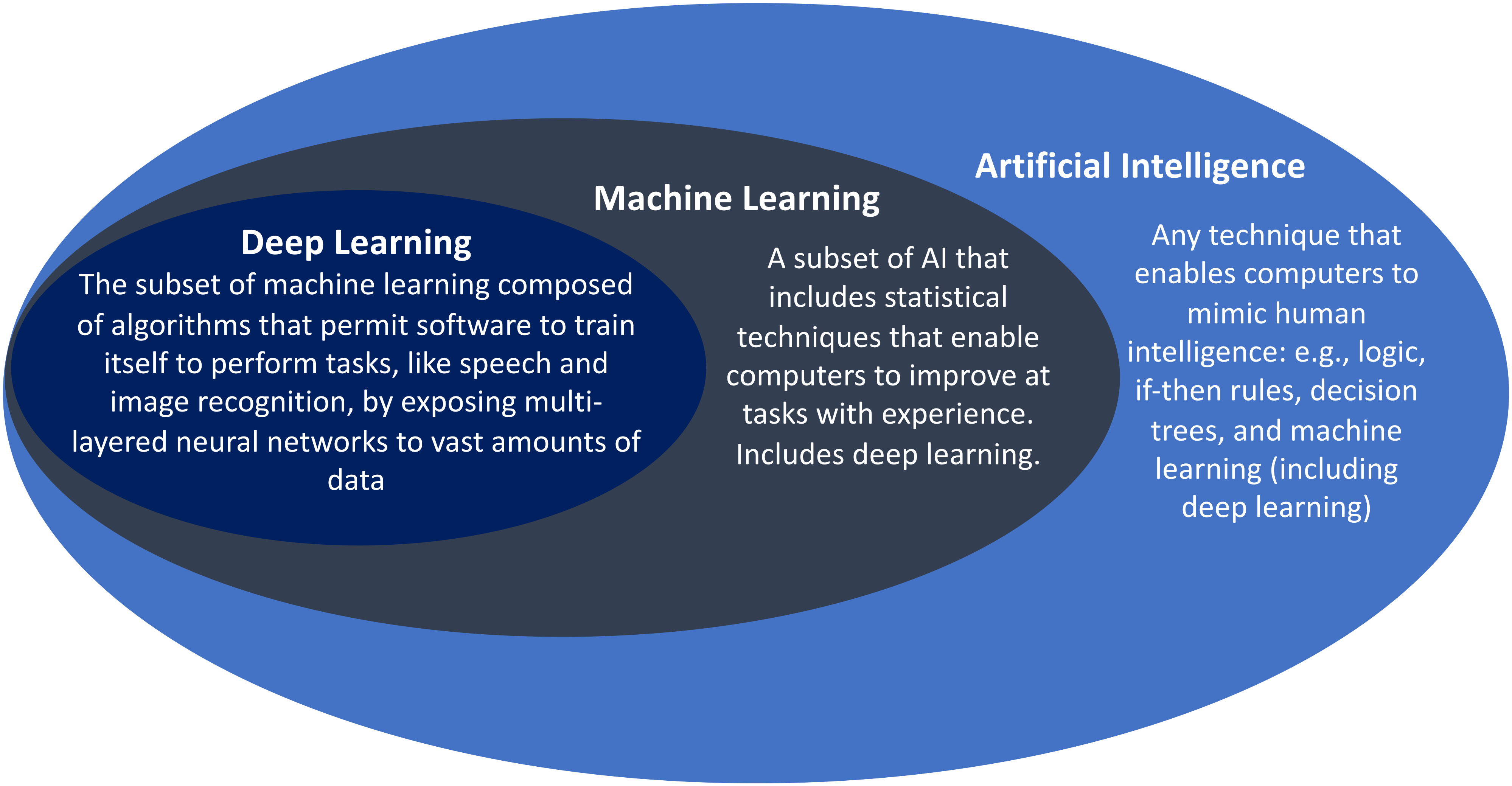
Artificial Intelligence

Any technique that enables computers to mimic human intelligence: e.g., logic, if-then rules, decision trees, and machine learning (including deep learning)

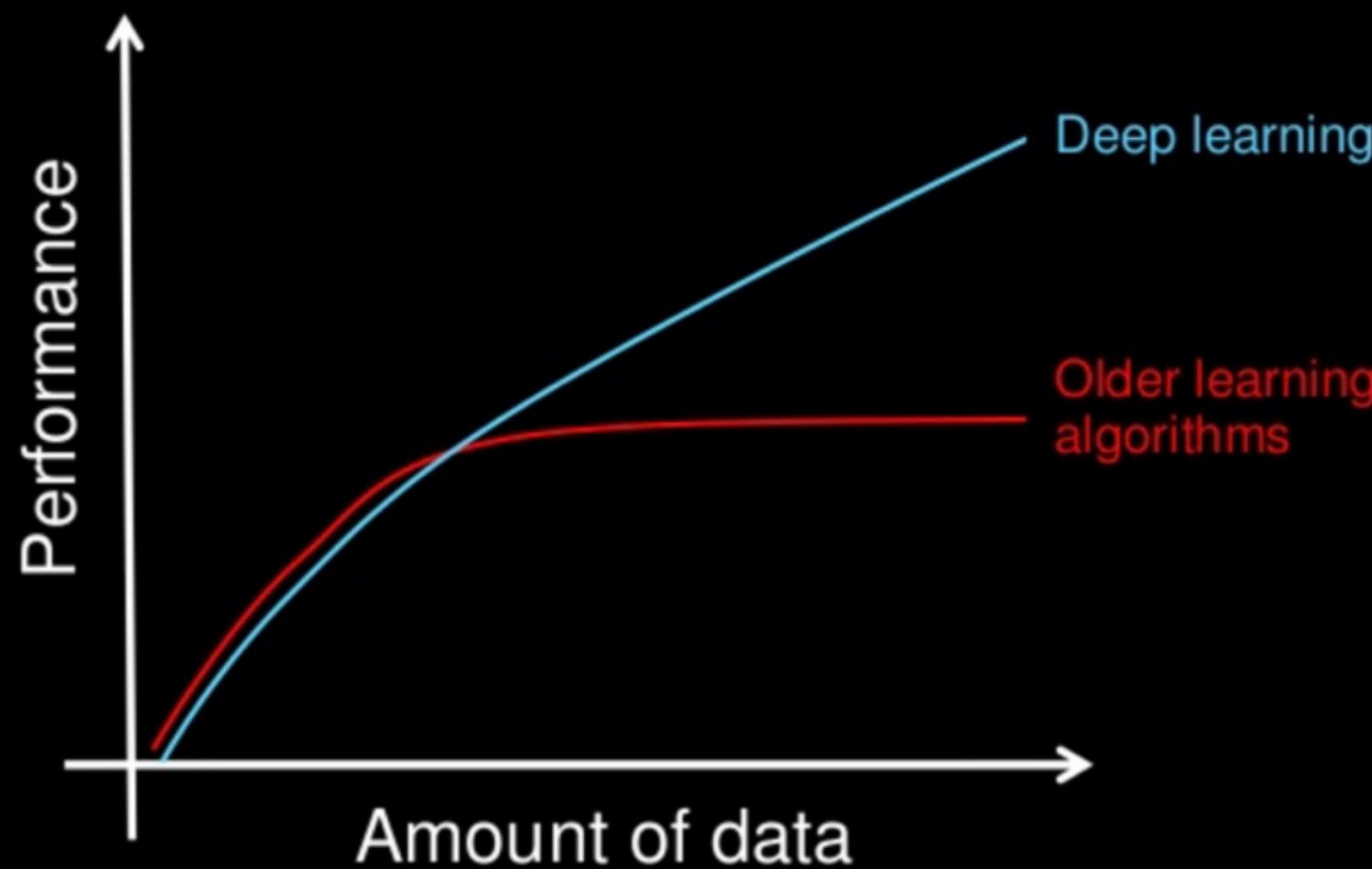
Briefly: What is AI?



Briefly: What is AI?



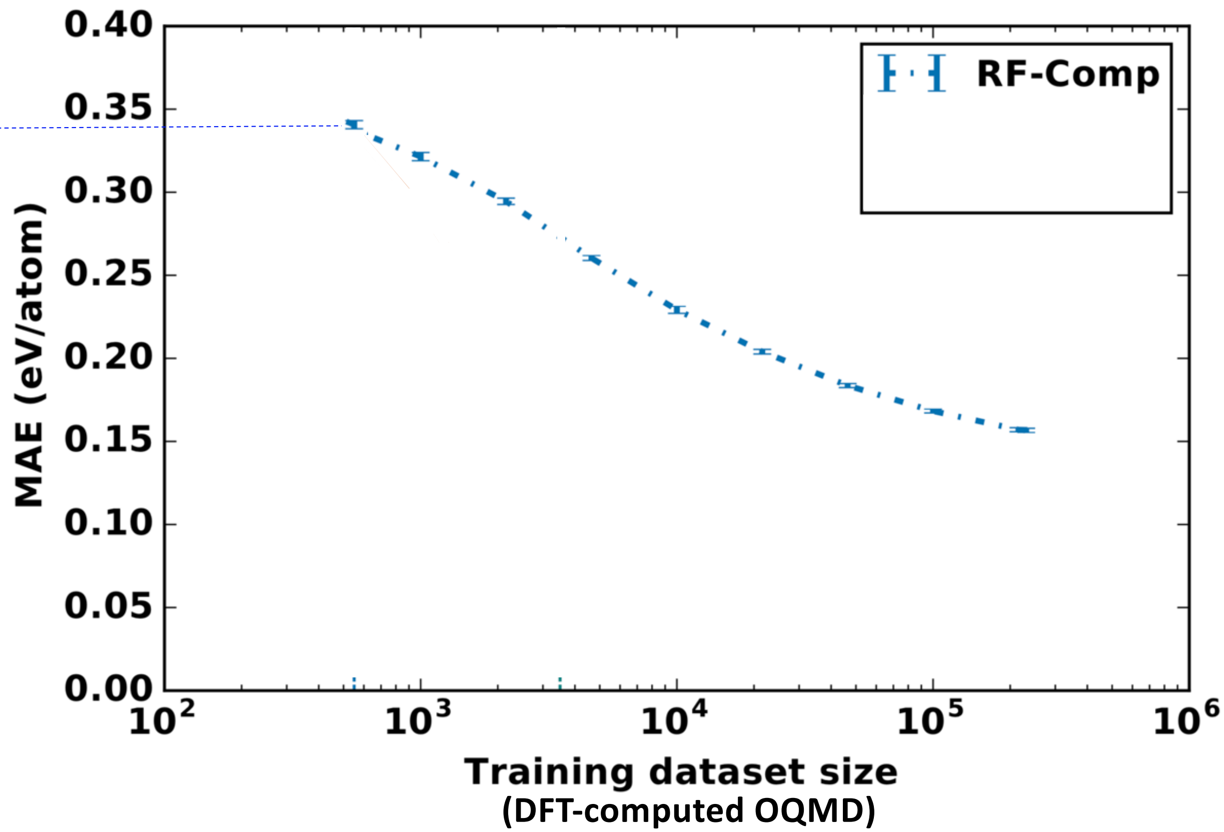
Why deep learning



Predicting formation enthalpies of crystalline materials

Best conventional machine learning method,
Random Forest:

a) Only elemental compositions



Given ΔH_f e.g.:
Cr_2Ni_3
Al_2O_3

Predict:
$TiO_2 \rightarrow ?$

Logan Ward et al., Phys Rev B, 2017

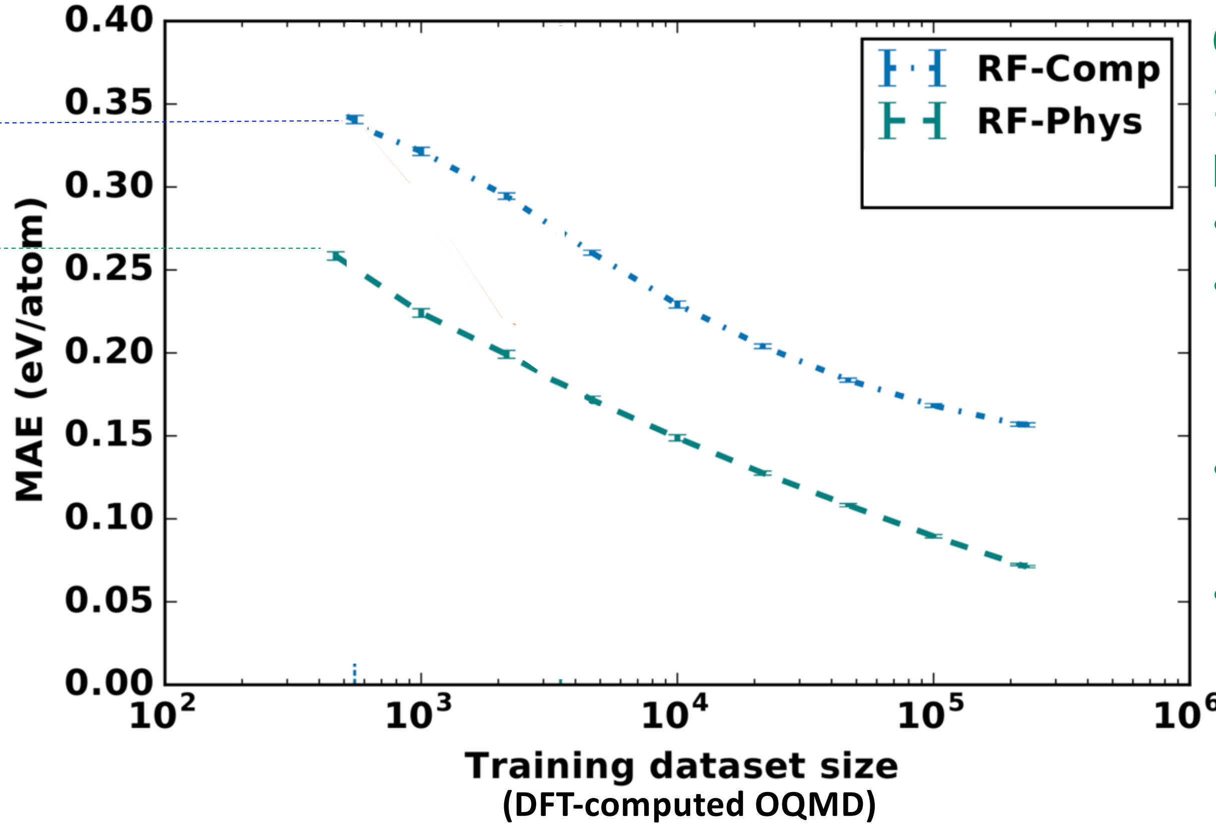
Predicting formation enthalpies of crystalline materials

Best conventional machine learning method, Random Forest:

a) Only elemental compositions

b) Also physical attributes

Logan Ward et al., Phys Rev B, 2017



Compute 145 physical properties:

- Stoichiometric
- Elemental property statistics
- Electronic structure
- Ionic compound

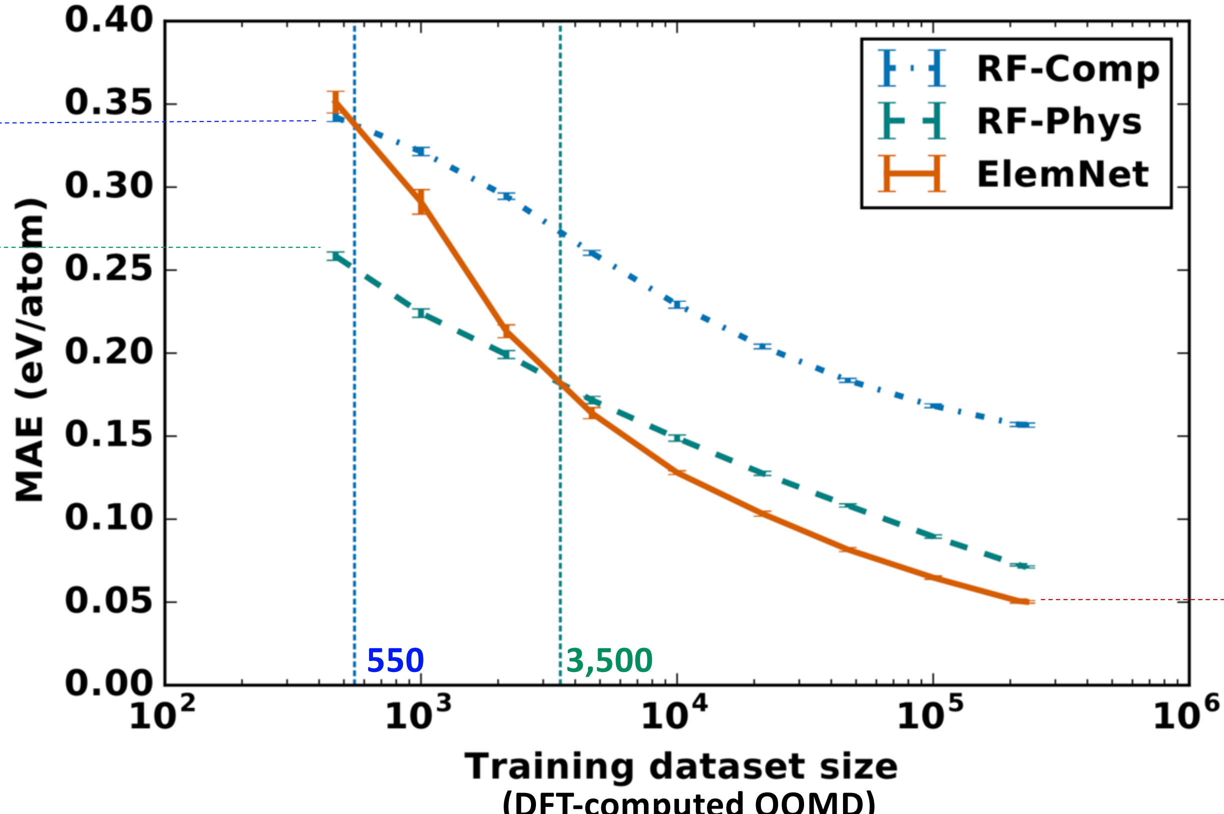
Predicting formation enthalpies of crystalline materials

Best conventional machine learning method,
Random Forest:

a) **Only elemental compositions**

b) **Also physical attributes**

Logan Ward et al., Phys Rev B, 2017



Dipendra Jha

ElemNet,
17-layer DNN:
Only elemental compositions

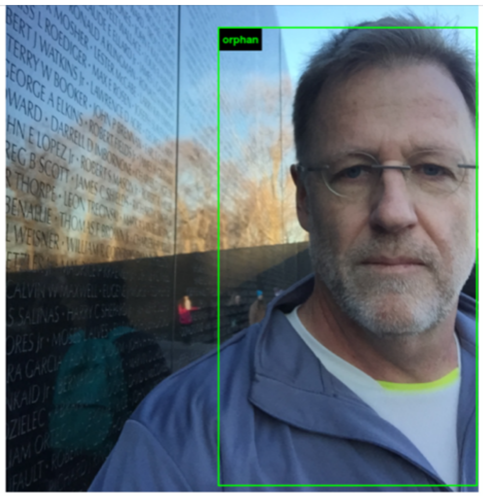
(Also runs 100x faster than RF)

Jha, Ward, et al., 2018.

New methods have limitations: E.g., ImageNet Roulette

judge, justice, jurist: *a public official authorized to decide questions brought before a court of justice*

- person, individual, someone, somebody, mortal, soul > adjudicator > judge, justice, jurist
- person, individual, someone, somebody, mortal, soul > worker > skilled worker, trained worker, skilled workman > official, functionary > judge, justice, jurist

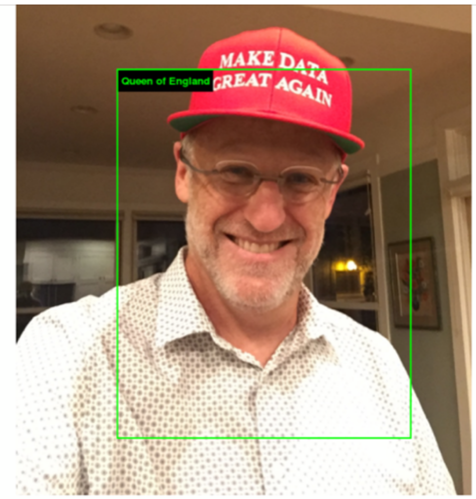


orphan: *a child who has lost both parents*

- person, individual, someone, somebody, mortal, soul > juvenile, juvenile person > child, kid, youngster, minor, shaver, nipper, small fry, tiddler, tike, tyke, fry, nestling > orphan

swami: *a Hindu religious teacher; used as a title of respect*

- person, individual, someone, somebody, mortal, soul > religious person > Hindu, Hindoo > swami



Queen of England: *the sovereign ruler of England*

- person, individual, someone, somebody, mortal, soul > leader > aristocrat, blue blood, patrician > female aristocrat > queen, queen regnant, female monarch > Queen of England

https://www.excavating.ai

Why are we excited about "AI for science"?

Push

- Step changes in AI/ML methods, notably deep neural networks
- Major advances in areas like machine translation, speech recognition, image processing
- New hardware specialized for deep neural networks

Pull

- Exploding volumes of data due to new sensors and instrumentation exceed human capabilities
- End of Moore's Law puts hard problems out of reach
- Growing complexity of science and engineering problems slowing rate of discovery

Argonne's major Initiatives offer compelling targets

Five major initiatives build on our core capabilities to deliver cutting-edge science and enable future energy technologies

"AI" is an opportunity in all five areas

Hard x-ray sciences

Transform understanding of materials and chemical systems through 3D microscopy

Advanced computing

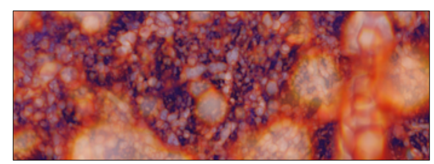
Deploy exascale computer and advance machine learning and quantum and neuromorphic computing

Materials and chemistry

Discover emergent phenomena and synthesize novel materials and chemical systems

The universe as our laboratory

Make leading contributions to physics experiments that explore the early universe and its dynamics



Energy manufacturing science and engineering

Create science-based approaches to speed scaling of manufacturing processes for energy technology

Things we can do 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 model real data

Example: synthetic sky in cosmology

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 $\Rightarrow$ end-to-end automated science

Create an ability to use AI for generating new theories that address problematical areas of existing theories

In 10 years …

- **Learned models begin to replace data**
  – Queryable, portable, pluggable, chainable, secure
- **Experimental discovery processes are dramatically refactored**
  – Models replace experiments, experiments improve models
- **Many questions are pursued semi-autonomously 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

http://bit.ly/2FOnJi3

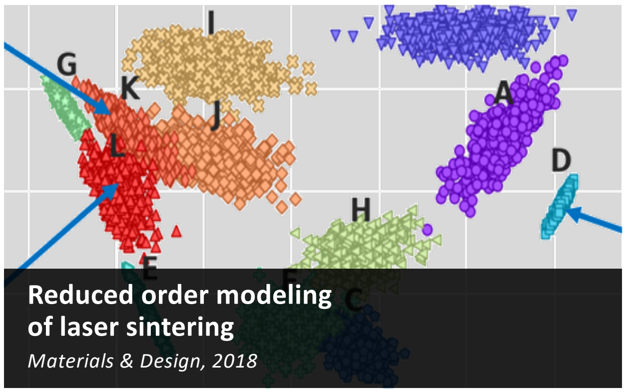
AI for Science @ Argonne targets both research and infrastructure

Applications	AI applications across science and engineering. Transformative approaches to simulation and experimental science.
Learning systems	AI software. Software infrastructure for managing data, models, workflows etc., and for delivering AI capabilities to 1000s of scientists and engineers.
Foundations	Mathematics, algorithms; general AI, reinforcement learning, uncertainty quantification, explainability, etc.
Hardware	Advanced hardware to support AI. Evaluation of new architectures and systems; exploration of neuromorphic and quantum as long term accelerators for AI.

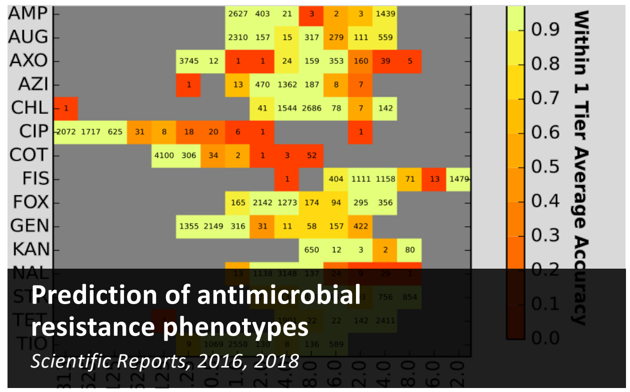
AI for Science @ Argonne targets both research and infrastructure

Applications	AI applications across science and engineering. Transformative approaches to simulation and experimental science.
Learning systems	AI software. Software infrastructure for managing data, models, workflows etc., and for delivering AI capabilities to 1000s of scientists and engineers.
Foundations	Mathematics, algorithms; general AI, reinforcement learning, uncertainty quantification, explainability, etc.
Hardware	Advanced hardware to support AI. Evaluation of new architectures and systems; exploration of neuromorphic and quantum as long term accelerators for AI.

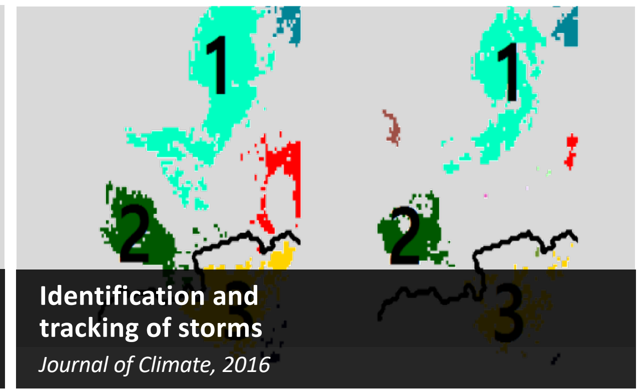
AI at Argonne: Dozens of applications projects



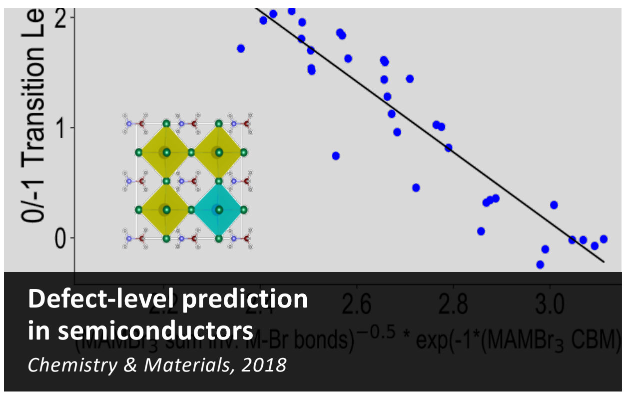
Reduced order modeling of laser sintering
Materials & Design, 2018



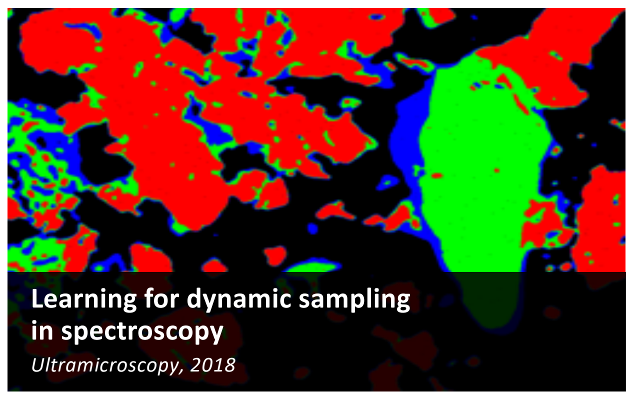
Prediction of antimicrobial resistance phenotypes
Scientific Reports, 2016, 2018



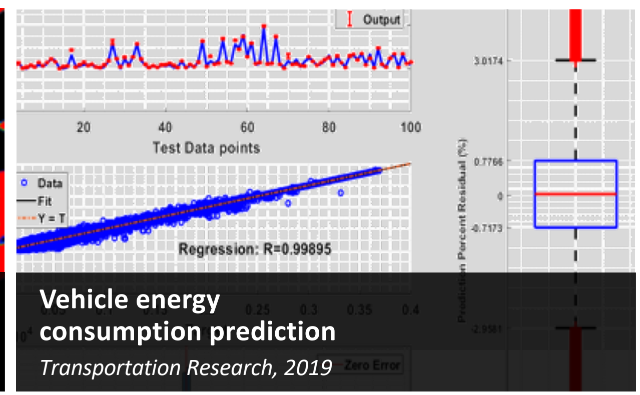
Identification and tracking of storms
Journal of Climate, 2016



Defect-level prediction in semiconductors
Chemistry & Materials, 2018



Learning for dynamic sampling in spectroscopy
Ultramicroscopy, 2018



Vehicle energy consumption prediction
Transportation Research, 2019

AI for Science: Advanced Photon Source Upgrade

AI can drive the scientific and measurement motifs enabled by APS-U

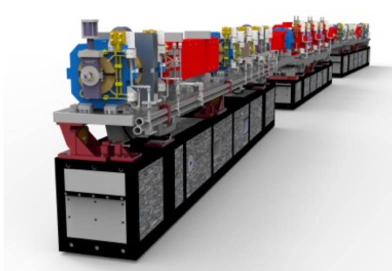
Detect rare events/features in large volumes with nanoscale resolution	Metal fatigue, solid-state batteries, brain circuitry
Capture dynamic processes	Catalyst coarsening, precision synthesis, additive manufacturing
Enable multidimensional inquiry, exploring spaces of higher dimension and size	High-entropy alloys, metal fatigue, catalysts

APS-U's 2-3 orders of magnitude increased brightness and coherent flux, will lead to:

- Too much data for humans to handle
- Data rates too fast for human management

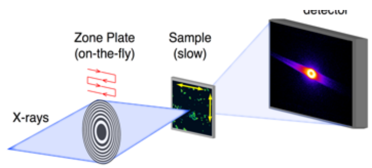
Control: real-time autonomous execution

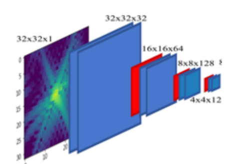
Analyze: reconstruct, feature extraction, viz, optimized photon dose



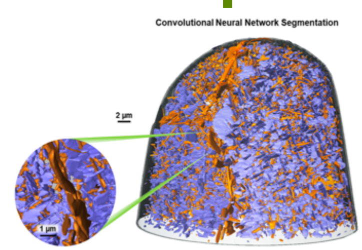
Advanced accelerator control:
100s of control points, 1000s of inputs

Fast, many megapixel detectors





AI at the edge:
autonomous data reduction near the source



Beyond the depth of focus limit: APS and ALCF

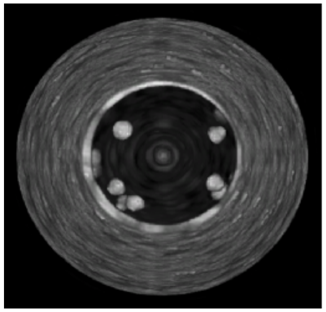
Depth of focus goes like DOF≃5.4(transverse resolution)^2/λ.

- Example: DOF=4 μm for 10 nm resolution at 10 keV
- We can't treat images as simple projections of an object when going beyond DOF!
- How to combine penetrating power of X rays with tomorrow's resolution limit?

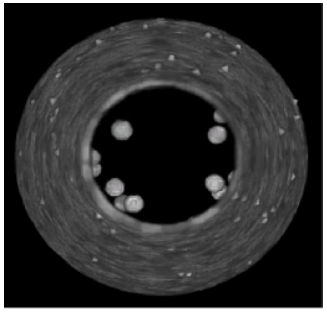
- **We can describe the forward problem using multislice wave propagation** [1,2]
- **We can then "learn" how to adjust a 3D object to agree with microscope measurements.**
  - First demo required 200,000 core hours at Argonne's LCRC [3]
- **Now using Automatic Differentiation (part of AI toolkits) to "learn" how to adjust a 3D object with additional info included**
  - AD applied to parallelized 2D ptychography [4]
  - Presently using Argonne's ALCF Data Science Program allocation, plus support from LDRD and also NIH
  - Ming Du, Saugat Kandel, Sajid Ali, Northwestern; ANL APS; ANL MCS; ANL MSD.

Simulations: 12 DOF thick

DOF effects ignored: reduced resolution



DOF effects included with multislice/AD: full resolution 3D imaging!



[1] Cowley, *Acta Cryst.* **10**, 609 (1957). [2] Li *et al.*, *Optics Express* **25**, 1831 (2017). [3] Gilles *et al.*, *Optica* **5**, 1078 (2018). [4] Nashed *et al.*, *Procedia Computer Science* **108**, 404 (2017)

AI in materials and chemistry

AI will accelerate discovery through autonomy in data collection and physics-based model building

Extracting rules from multimodal data

Discover New Phenomena in Heterogeneous Data Streams

Supervised and unsupervised learning will uncover pathways and intermediates encoded in multidimensional data

Predicting transformations across scales

Achieve Deterministic Control of Multiscale Transformations

Active learning and deep neural networks will navigate reaction networks and multicomponent phase behavior

Translating design to creation

Deliver on the Promise of Predictive Design of Matter

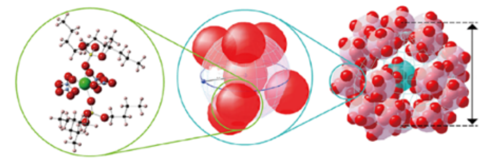
Reinforcement learning and decision trees will autonomously explore and control multiscale synthesis

Data Needs

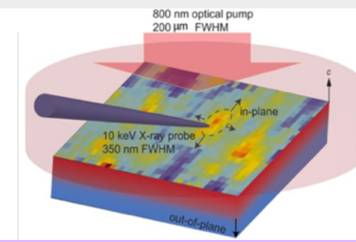
- Multimodal Streams at APS (10^4-10^5/day)
- High Throughput Synthesis (10^3/day)
- Simulation/Synthetic Data (10^1-10^4/day)

AI Needs

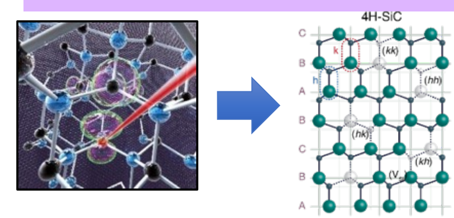
- Advanced sampling for sparse data
- Rare-event detection and interpretation
- Domain specific models



Interrogate organization, dynamics, and reactivity of multiscale systems



Derive rules of dynamical organization among charges, spins, and atoms



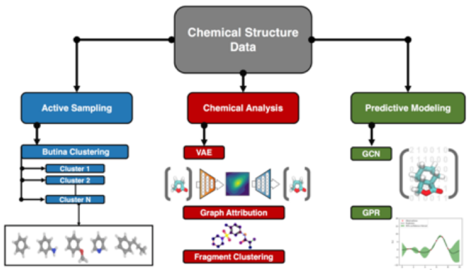
Design and synthesize functional atomic and defect landscapes

Opportunities for AI in Materials and Chemistry: Findings of an Argonne internal task force

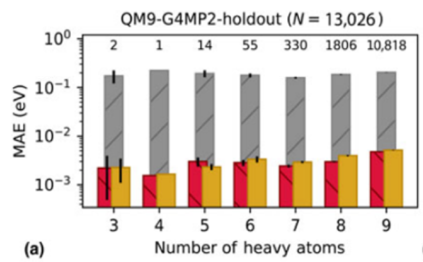
§	Area	AI requirements and challenges
5.1	AI-Accelerated Ab Initio Molecular Dynamics for Catalysis	Methods development to enable application of ML/AI methods to extremely large collections of samples obtained from simulation studies, and for efficient coupling of simulation and AI components.
5.2	Ultra-Fast Simulations of Complex Materials	Processing billions of DFT energy evaluations is likely to require extremely large neural networks. Handling data from multiple sources is also a key need.
5.3	Designing New Chemical Pathways Automatically	Tight integration with experiment. Reinforcement learning and active learning algorithms to guide experimental campaigns. Representation and update of kinetic table and associated uncertainties.
5.4	Real-time Inversion of Multi-modal Characterization Data	Requires methods for integrating physical constraints into neural networks (NNs). May also build up large enough NNs to require specialized AI accelerators.
5.5	Panoramic Synthesis for Discovery and Deployment of New Materials	Would benefit from symbolic AI to create human-interpretable (and, ideally, scientifically testable) design rules for panoramic synthesis.
5.6	AI-Driven Material Discovery for Energy Storage	Tight integration with computational simulation. Reinforcement learning and active learning algorithms to guide computational campaigns.
5.7	Discovery and Design of Magnetic Topological Materials and Magnetic Order	Learning from small data. Transfer learning between different classes of materials. Integration of experimental and simulation data.
5.8	AI-Generated Designs of Unconventional Structures	Requires method development for generative models for networks/paths and supervised learning methods on graphs/path data.
5.9	Comprehensive Atlas of Phase Diagrams of All (Meta)Stable Materials	Requires advances in natural language processing (NLP) and in methods for propagating uncertainty through many different supervised learning and physical models.
5.10	Optimizing Gas-phase Chemistry for Scale-up of Complex Materials	AI-based surrogate models for manufacturing processes are needed that can enable near-real-time feedback; current multi-scale simulation methods take days or weeks.

AI for materials and chemistry: Examples

Molecular Informatics

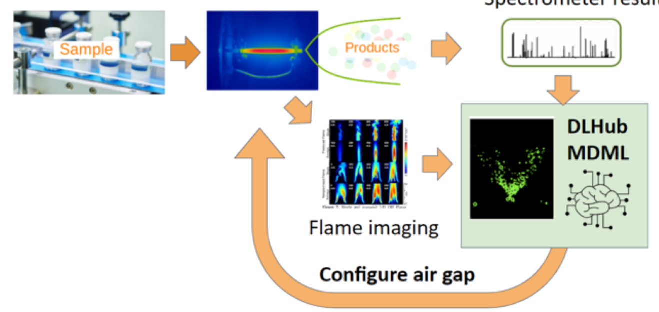


Sivaraman, Jackson et al. (2019)



Ward *et al.* (2019)

AI in Manufacturing



AI-Enhanced Imaging

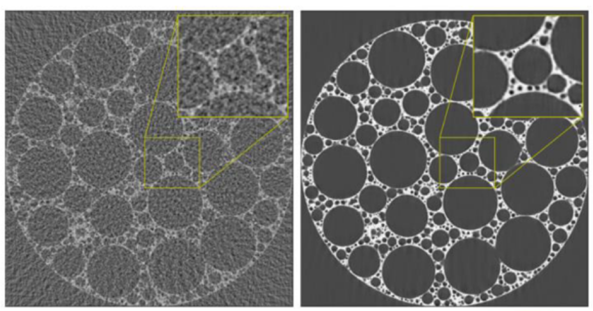
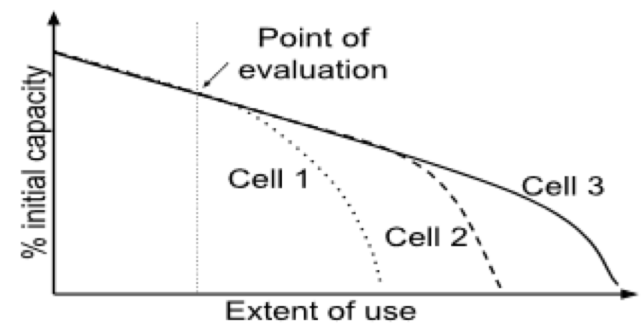


Image produced by conventional reconstruction.

Image following enhancement by TomoGAN.

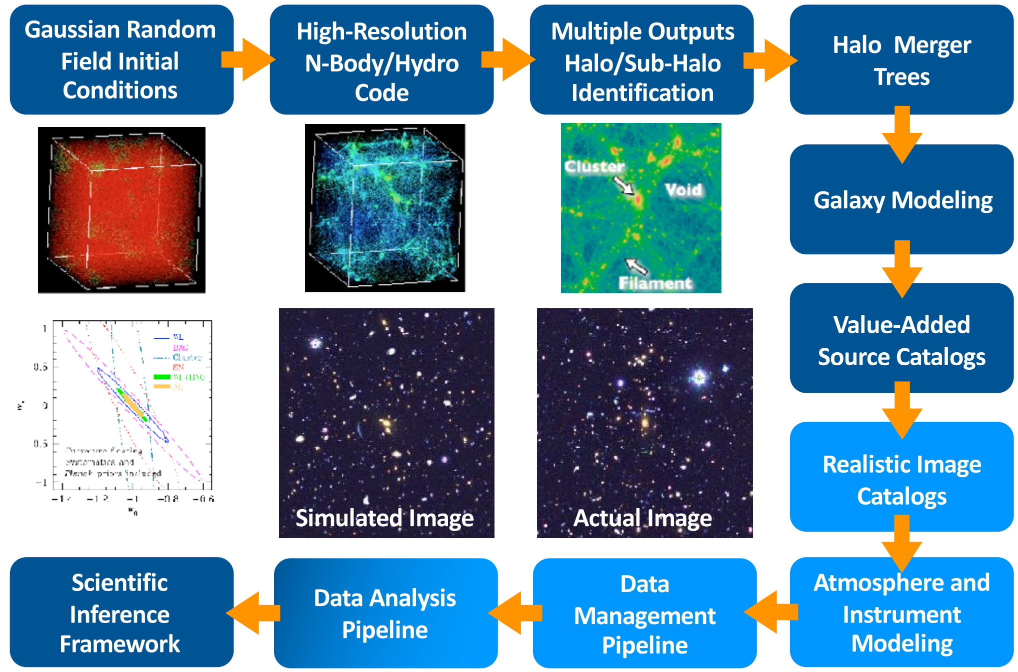
Liu et al. (2019)

Battery Life Prediction



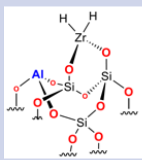
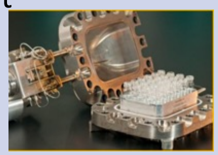
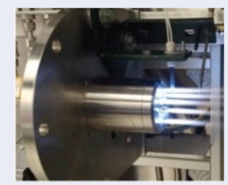
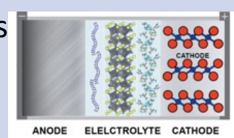
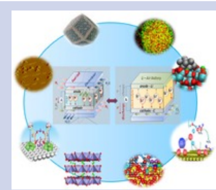
AI for science: High Energy Physics at Argonne

- **Energy/Intensity Frontier:**
  - GANs for detector simulation
  - CNN-based event classification
  - Learn detector effects that impact physics
  - Hyperparameter scans to reduce reconstruction time
  - Anomaly detection with autoencoders
  - Neuromorphic models/accelerators for use in triggers
- **Cosmic Frontier (SciDAC-4):**
  - GANs for object catalog emulation
  - GP/DL-based emulators for diverse cosmic probe summary statistics
  - CNN-based image classification
  - AI/ML-based methods for large-scale training data production and estimation methodology for photometric redshifts
  - Likelihood emulation and likelihood-free methods for statistical inference



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 reshift estimation, 5) Likelihood-free methods for inference *[Work performed under the Argonne-led SciDAC-4 project: "Inference and Machine Learning at Extreme Scales"]*

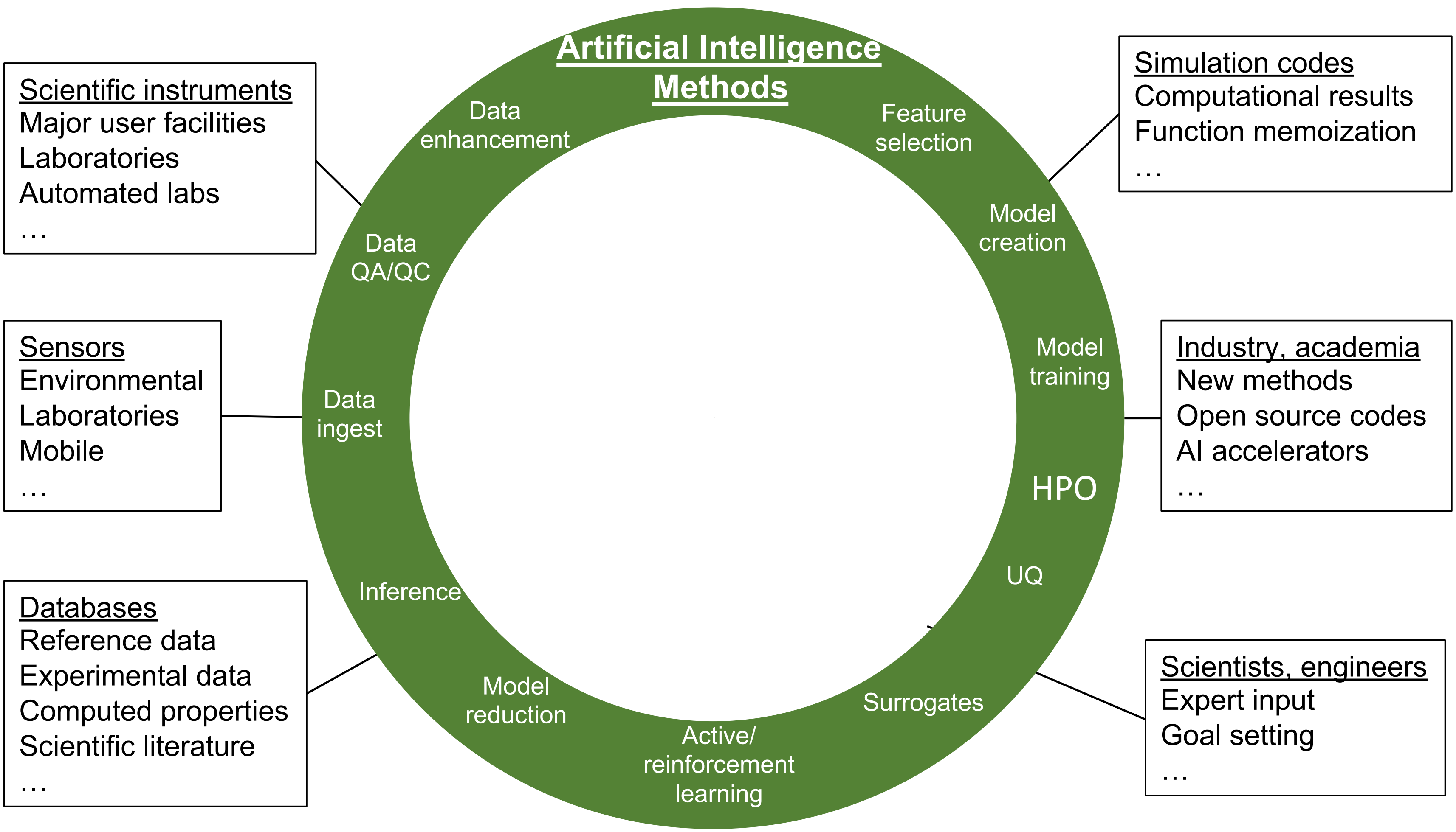
AI for transformative manufacturing science

Science Impact	Available Data	AI Role	Manufacturing Impact
Mimic Pt-group metal catalysis using earth-abundant elements 	High throughput catalyst performance testing 	Data simulation and prediction	Predict economical catalytic conversions
Expand thermochemical tables ATcT	Rotational and mass spectra of gas-phase reaction products	On-the-fly spectral deconvolution, queryable models for pathway prediction	Predict new synthesis pathways via flame spray pyrolysis 
Improve models of ion transport in materials 	Real-time battery performance and lifetime data	Anomaly detection, correlation modeling 	Improve battery performance

AI for Science @ Argonne targets both research and infrastructure

Applications	AI applications across science and engineering. Transformative approaches to simulation and experimental science.
Learning systems	AI software. Software infrastructure for managing data, models, workflows etc., and for delivering AI capabilities to 1,000s of scientists and engineers.
Foundations	Mathematics, algorithms; general AI, reinforcement learning, uncertainty quantification, explainability, etc.
Hardware	Advanced hardware to support AI. Evaluation of new architectures and systems; exploration of neuromorphic and quantum as long term accelerators for AI.

New applications require new methods

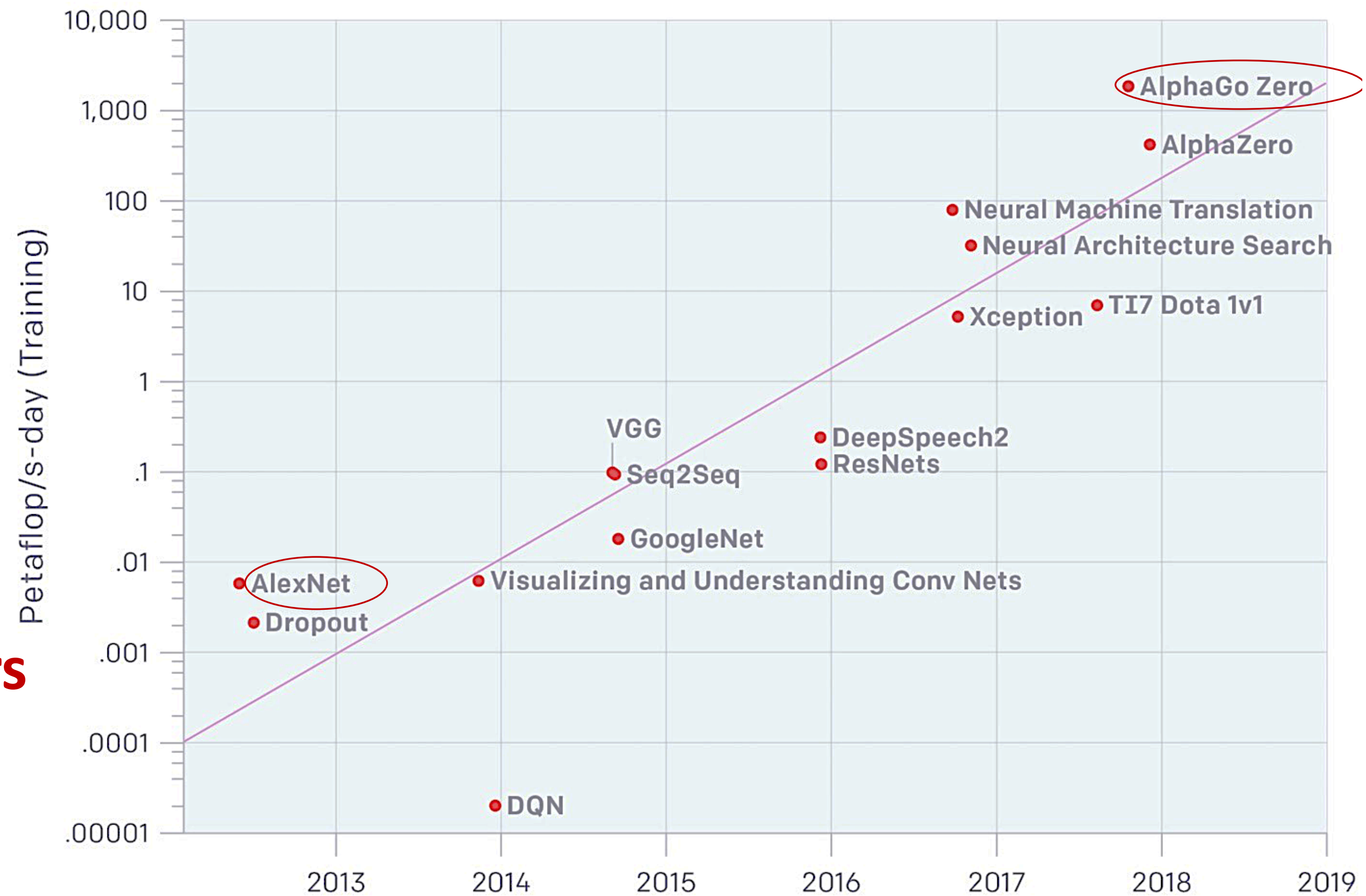


Deep learning uses massive computing

Exaflop/s-days used to train:

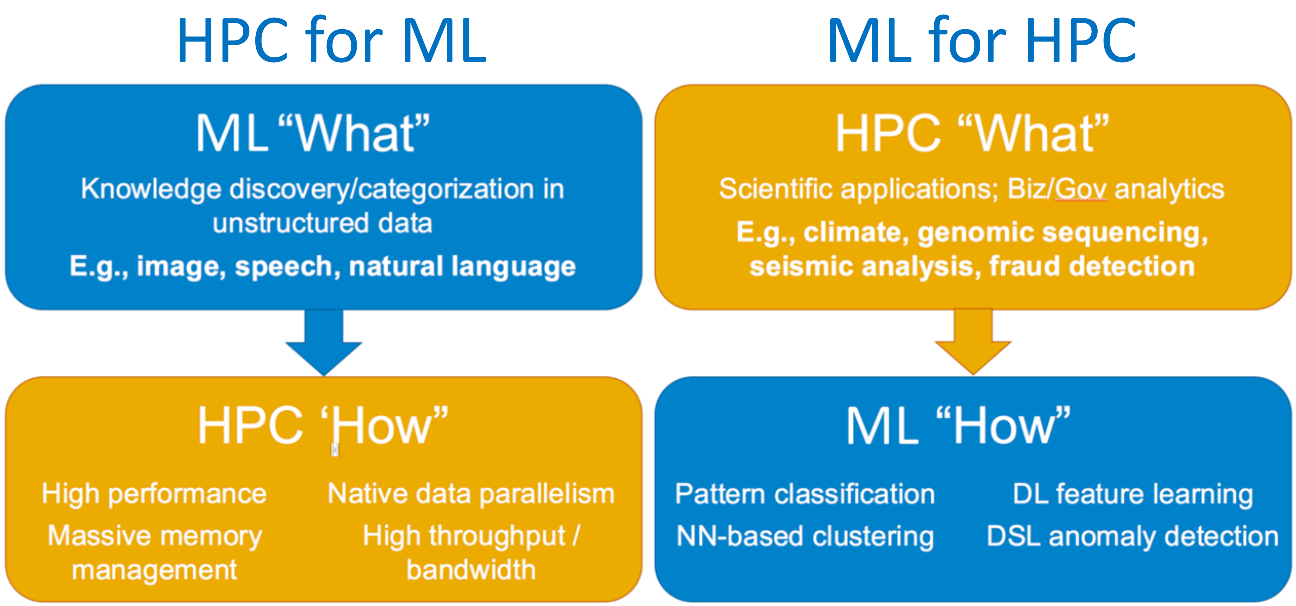
- AlexNet: 0.000007 (in 2012)
- AlphaGo Zero: 2 (in 2017)

x 300,000 in 5.5 years



Source: DeepAI

Relationships between AI/ML and HPC

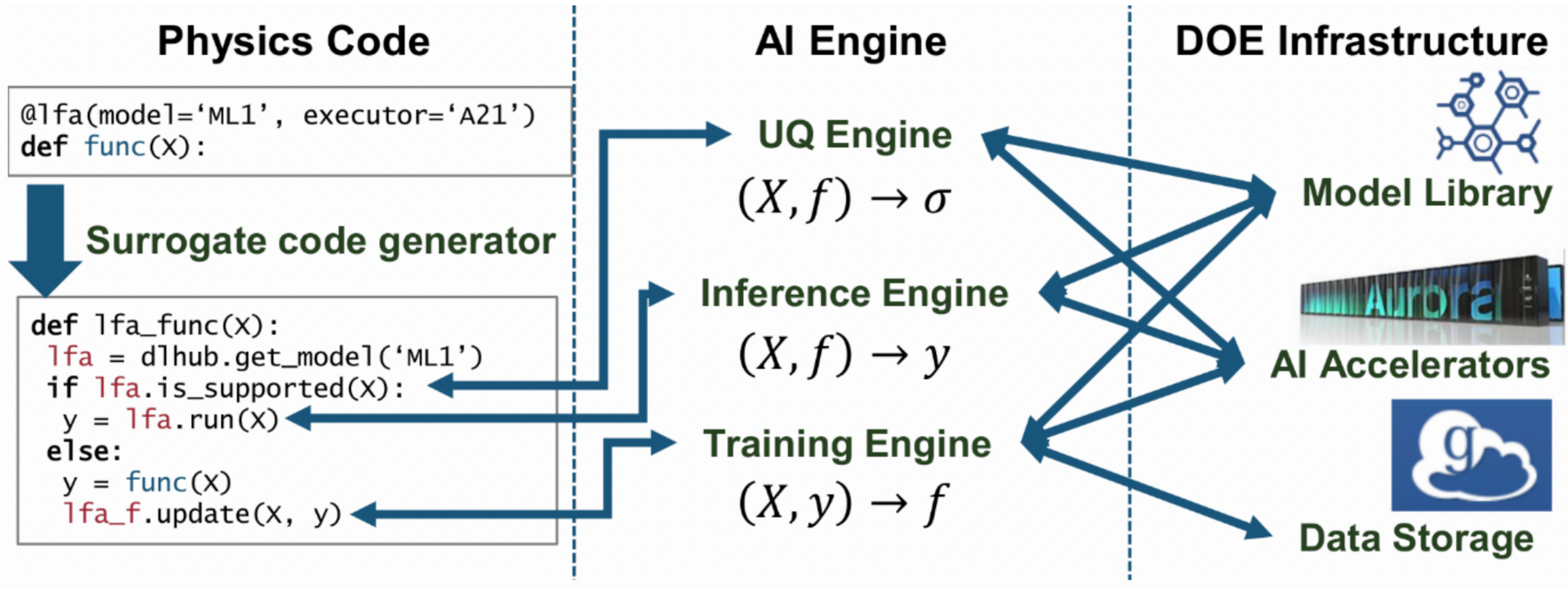


Ack: Gadi Singer

- HPC for AI/ML: HPC Technologies are applied to learning tasks to accelerate computation and/or solve larger problems
- AI/ML for HPC: Learning Technologies are applied to HPC computations to improve their performance in some way, e.g., by choosing the next simulation(s) to perform

Robust Learned Function Accelerators (RLFAs)

Fluidity between simulations and learned models

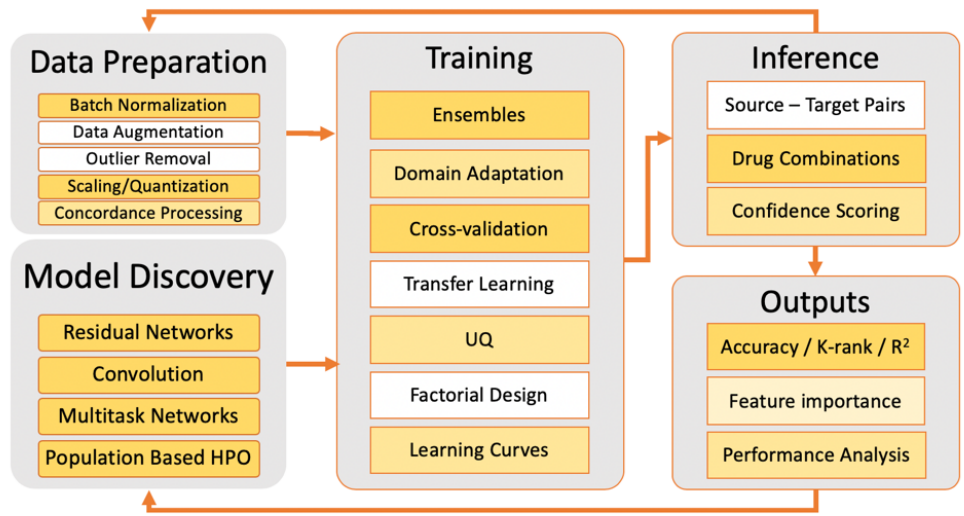


Logan Ward, Ben Blaiszik, et al.

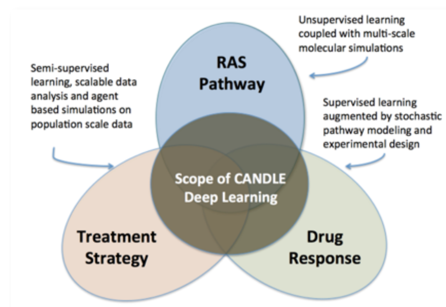
CANDLE: Exascale Deep Learning Tools

Deep Learning Needs Exascale

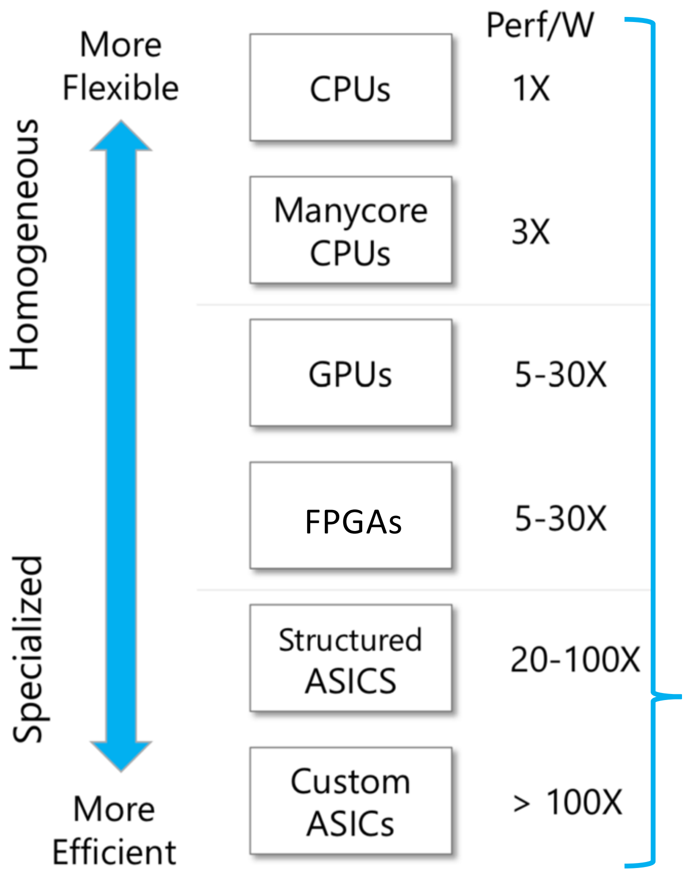
- Automated model discovery
- Hyper parameter optimization
- Uncertainty quantification
- Flexible ensembles
- Cross-Study model transfer
- Data augmentation
- Synthetic data generation
- Reinforcement learning



https://github.com/ECP-CANDLE

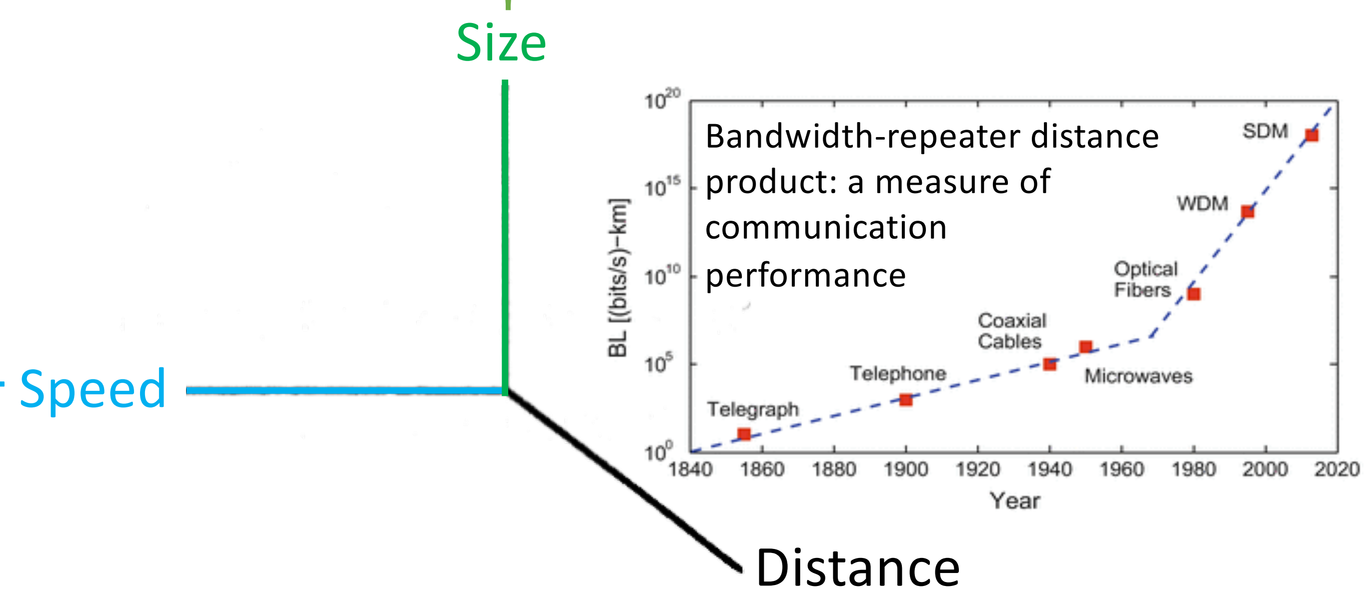


A rapidly evolving computing/data continuum



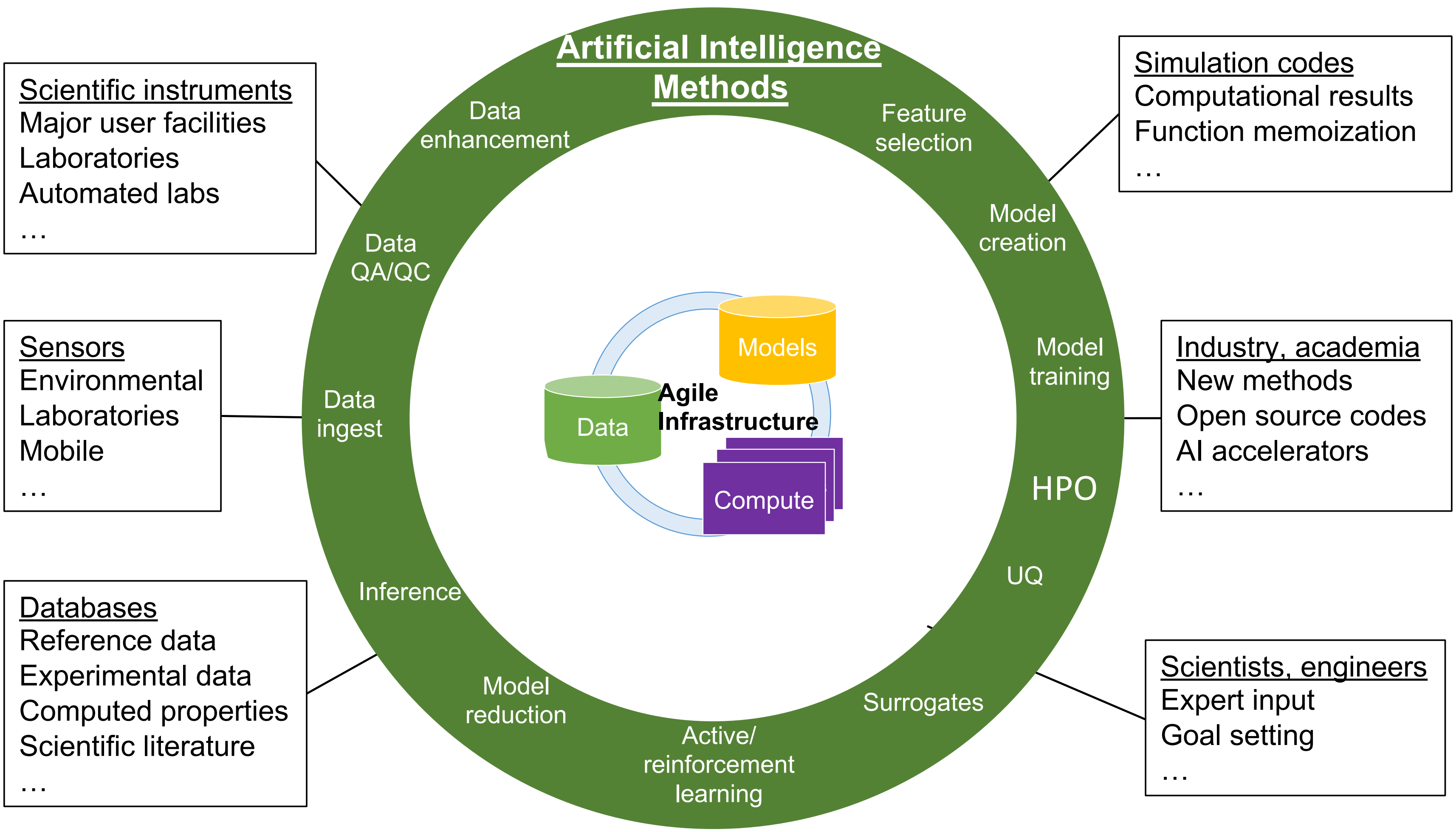
IoT/Edge — Fog — HPC/Cloud

Size	Nano	Micro	Milli	Server	Fog	Campus	Facility
Example	Adafruit Trinket	Particle.io Boron	Array of Things	Linux Box	Co-located Blades	1000-node cluster	Datacenter
Memory	0.5K	256K	8GB	32GB	256G	32TB	16PB
Network	BLE	WiFi/LTE	WiFi/LTE	1 GigE	10GigE	40GigE	N*100GigE
Cost	$5	$30	$600	$3K	$50K	$2M	$1000M



Sources: http://bit.ly/2SDGHzT, https://doi.org/10.1007/978-3-319-31903-2_8, Pete Beckman

New methods require new mechanisms



Project Celerity*

(My "Office of Science Distinguished Scientists Fellow" project)

Identity new mechanisms needed to bridge the gap between new (especially "AI-first") scientific applications and the emerging data/computing continuum

Via a process of experimentation, discussion, and debate

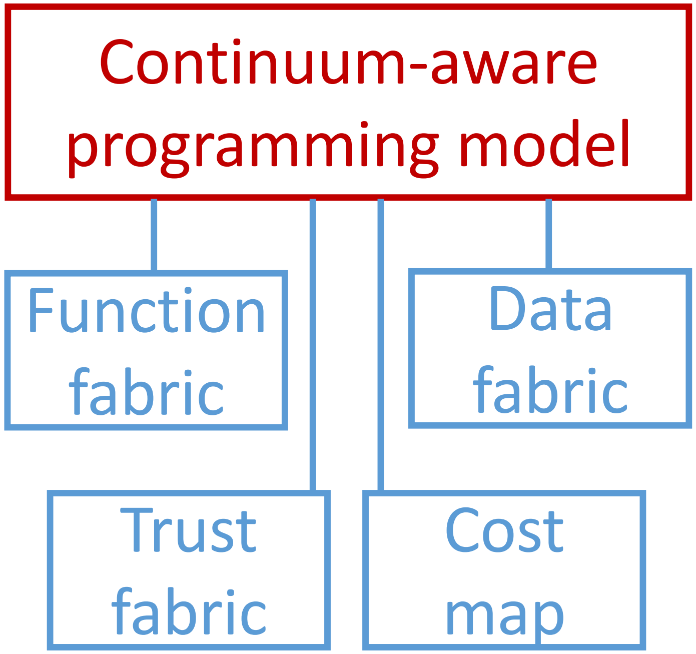
* **Celerity** (n) rapidity; swiftness; speed – from Latin celeritas, from which also **c** for speed of light in vacuum

Bridging the gap

What mechanisms will facilitate the programming of this distributed, heterogeneous, dynamically evolving AI-first continuum?

Some things that we surely need:

- **Function**: Compute wherever is fastest, cheapest, closest, most accurate, …
- **Data**: Access where fastest, cheapest, closest, most accurate, …
- **Trust**: Balance certainty vs. cost
- **Cost**: Useful estimates of the state of this dynamic system



Coding the continuum:
Elements of an open solution

globus labs

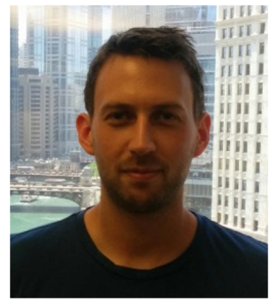
Thanks to colleagues, especially:


Rachana Ananthakrishnan


Yadu Babuji


Ben Blaiszik


Kyle Chard


Ryan Chard


Zhuozhao Li


Tyler Skluzacek


Steve Tuecke

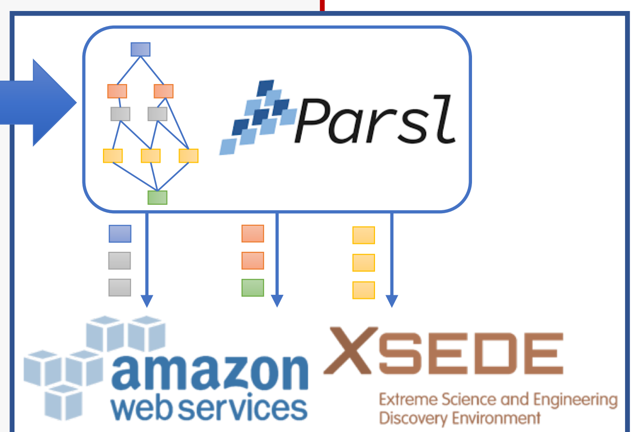

Anna Woodard


Logan Ward

Model registry	DLHub
Flows	Automate
Write programs	Parsl
Cost map	SCRIMP
Function fabric	funcX
Data fabric	Data services
Trust fabric	Auth

Coding the continuum: Elements of an open solution

```python
# App that estimates pi by placing points in a box
@python_app
def pi(total):
    import random

    # Set the size of the box (edge length) in which we drop random points
    edge_length = 10000
    center = edge_length / 2
    c2  = center ** 2
    count = 0

    for i in range(total):
        # Drop a random point in the box.
        x,y = random.randint(1, edge_length),random.randint(1, edge_length)
        # Count points within the circle
        if (x-center)**2 + (y-center)**2 < c2:
            count += 1

    return (count*4/total)

# App that computes the average of the values
@python_app
def avg_points(a, b, c):
    return (a + b + c)/3

# Estimate three values for pi
a, b, c = pi(10**6), pi(10**6), pi(10**6)

# Compute the average of the three estimates
avg_pi  = avg_points(a, b, c)
```



https://arxiv.org/pdf/1905.02158

http://parsl-project.org

Model registry	DLHub
Flows	Automate
Write programs	**Parsl**
Cost map	SCRIMP
Function fabric	funcX
Data fabric	Data services
Trust fabric	Auth

Coding the continuum: Elements of an open solution

```
In [1]: from funcx_sdk.client import FuncXClient

        fxc = FuncXClient()

In [2]: func = """
        def add(data):
            sum_val = sum(data['data'])
            return sum_val
        """

In [3]: fxc.register_function("add_func", func, description="Sum a list of numbers.")

In [4]: input_data = [1, 2, 3]

        res = fxc.run(input_data, "user#laptop", "add_func")

In [5]: print(res)

        6
```

Portable code

Python
Docker, Shifter, Singularity

Any access

SSH, Globus, cluster or HPC scheduler

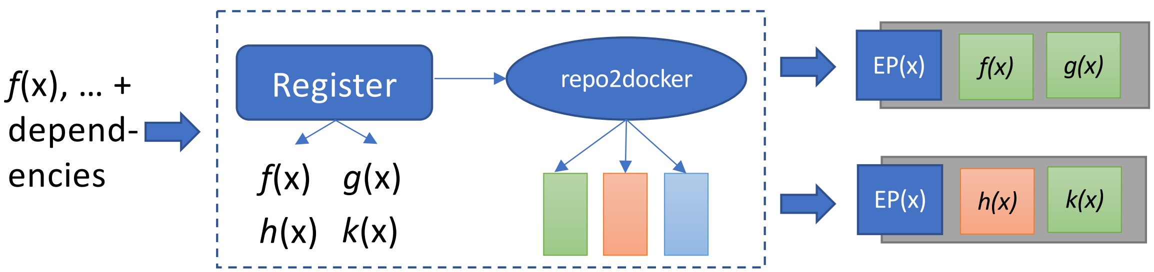
Any computer

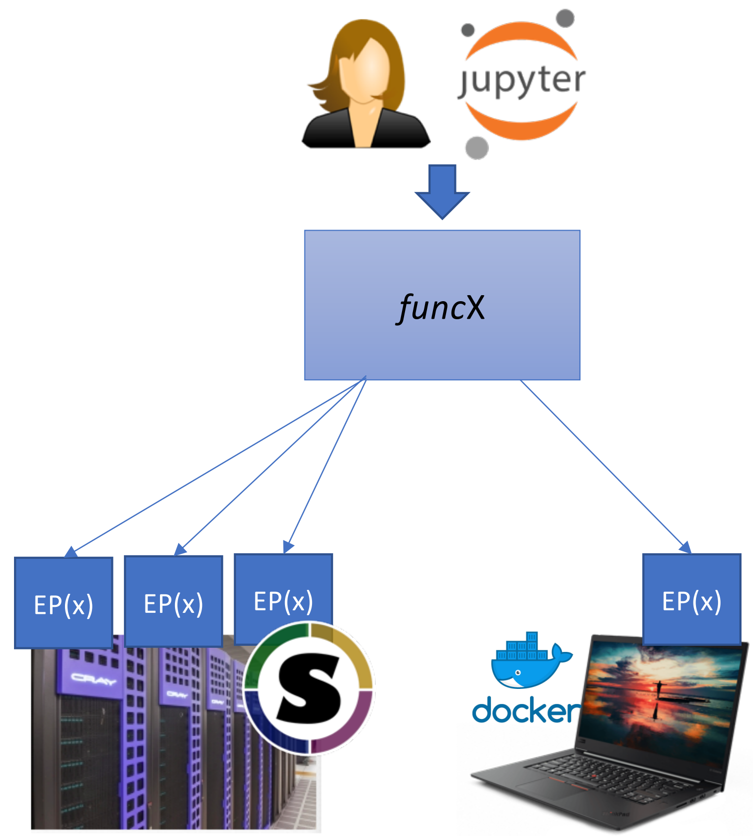
Clusters, clouds, HPC, accelerators

Model registry	DLHub
Flows	Automate
Write programs	Parsl
Cost map	SCRIMP
Function fabric	funcX
Data fabric	Data services
Trust fabric	Auth

funcX: Transform clouds, clusters, and supercomputers into high-performance function serving systems

```
In [1]: from funcx_sdk.client import FuncXClient

        fxc = FuncXClient()

In [2]: func = """
        def add(data):
            sum_val = sum(data['data'])
            return sum_val
        """

In [3]: fxc.register_function("add_func", func, description="Sum a list of numbers.")

In [4]: input_data = [1, 2, 3]

        res = fxc.run(input_data, "user#laptop", "add_func")

In [5]: print(res)

        6
```

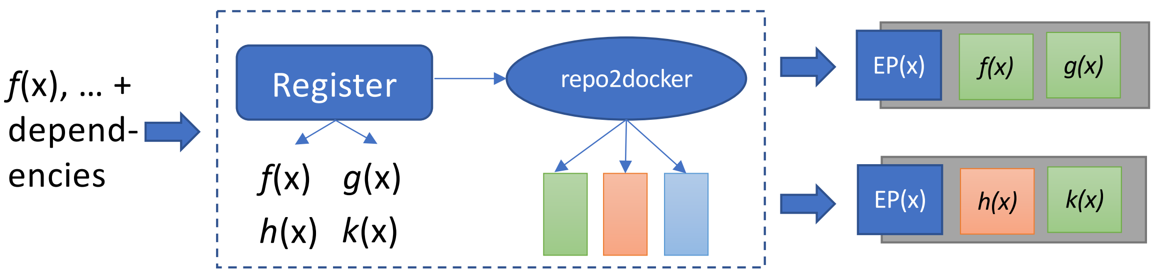


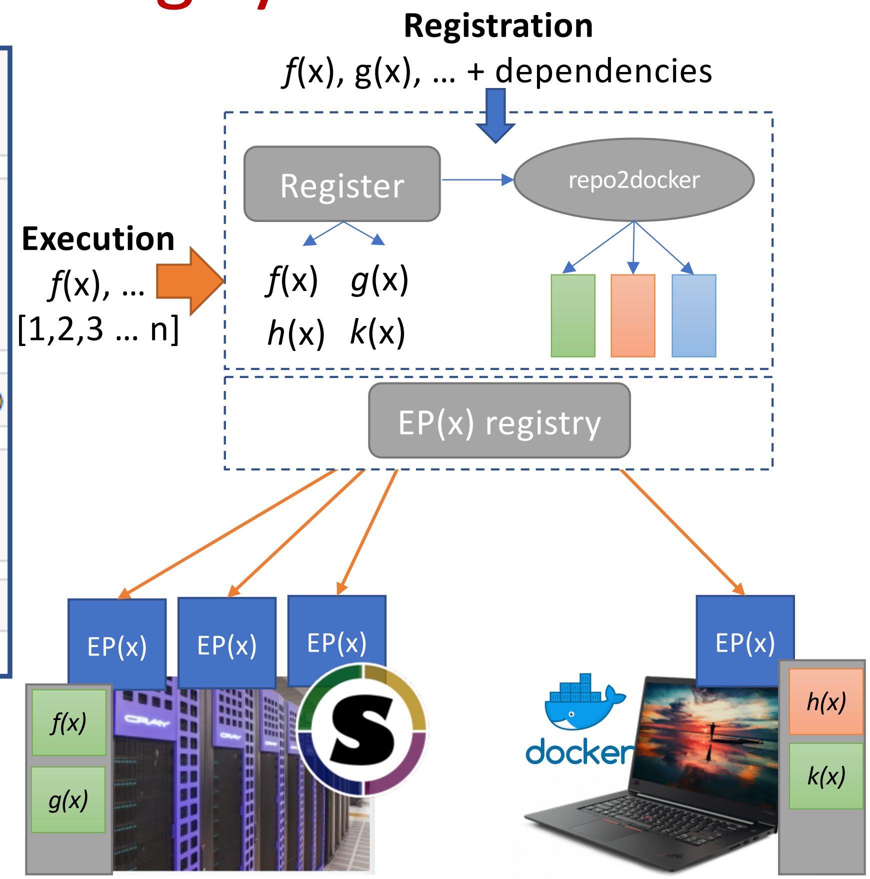


Simply deploy *funcX* endpoint to transform a computer into a function serving system

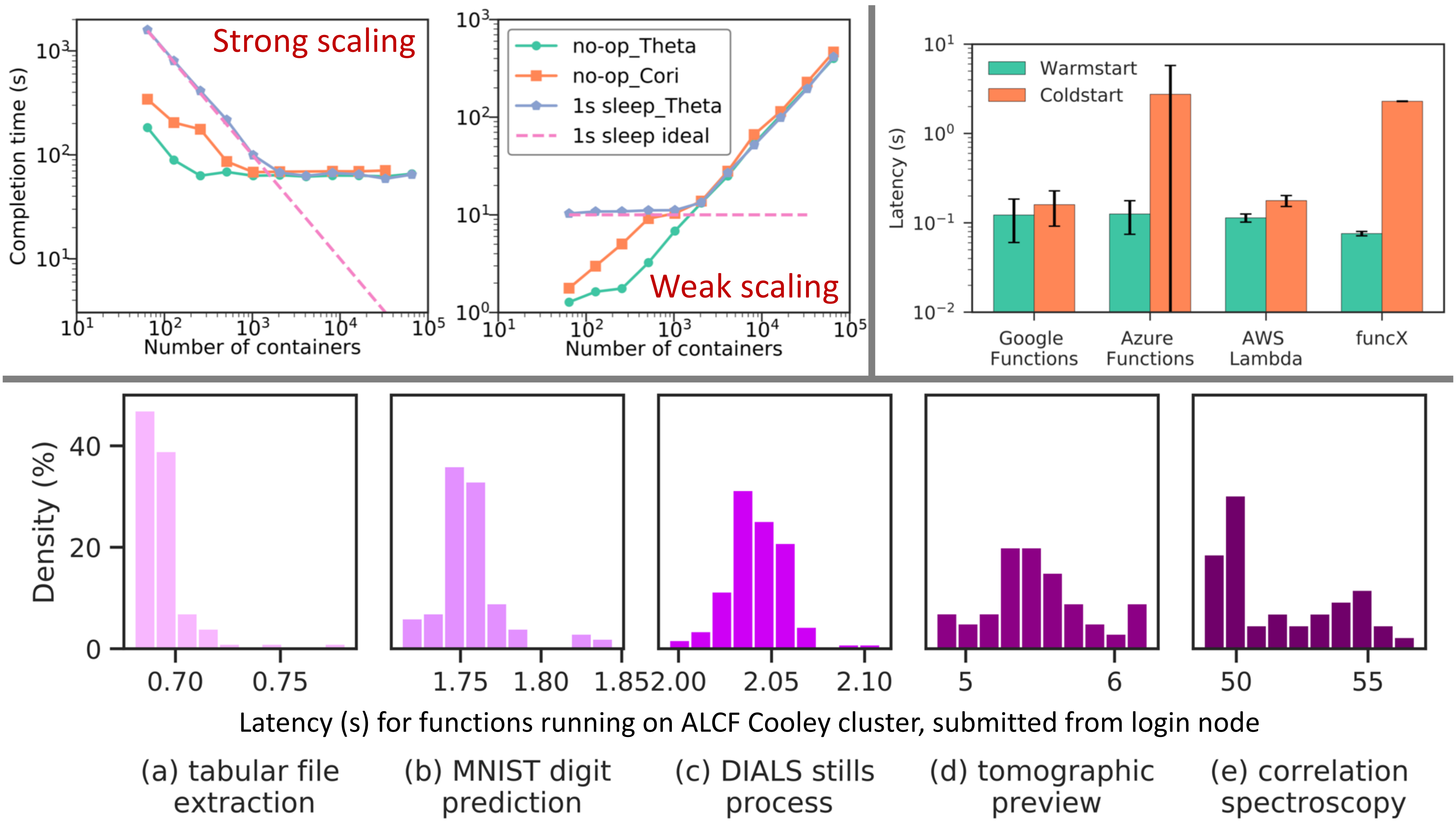
funcX: Transform clouds, clusters, and supercomputers into high-performance function serving systems

```
In [1]: from funcx_sdk.client import FuncXClient

        fxc = FuncXClient()

In [2]: func = """
        def add(data):
            sum_val = sum(data['data'])
            return sum_val
        """

In [3]: fxc.register_function("add_func", func, description="Sum a list of numbers.")

In [4]: input_data = [1, 2, 3]

        res = fxc.run(input_data, "user#laptop", "add_func")

In [5]: print(res)

        6
```





Simply deploy *funcX* endpoint to transform a computer into a function serving system


Strong scaling
Completion time (s)
Number of containers
no-op_Theta
no-op_Cori
1s sleep_Theta
1s sleep ideal
Weak scaling
Number of containers
Warmstart
Coldstart
Latency (s)
Google Functions
Azure Functions
AWS Lambda
funcX
Density (%)
Latency (s) for functions running on ALCF Cooley cluster, submitted from login node
(a) tabular file extraction
(b) MNIST digit prediction
(c) DIALS stills process
(d) tomographic preview
(e) correlation spectroscopy

DLHub: Organizing and Serving Models

DLHub
Data and Learning Hub for Science
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
- Citable DOIs for reproducible science

Argonne Advanced Computing LDRD

Models and Processing Logic as a Service

Describe
- Specify the model files
- Mark up the model with information to make it discoverable and usable

```
from dlhub_sdk.models.servables.keras import KerasModel
m = KerasModel.create_model("p1b1-example.h5")

m.set_title("CANDLE Pilot 1 - Benchmark 1")
m.set_name("candle_p1b1")
m.set_domains("genomics","biology","HPC")
```

Publish
- Register with DLHub for containerization as a servable
- DLHub service creates unique endpoint for servable

```
from dlhub_sdk.client import DLHubClient
dl = DLHubClient()
dl.publish(m)
```

Discover
- Discover servables with advanced search capabilities through Python SDK or web UI

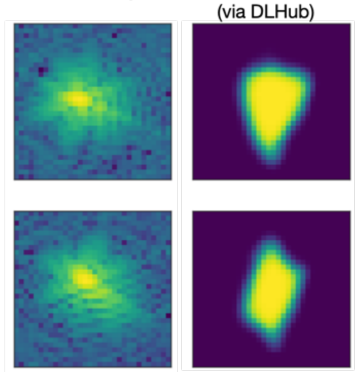
Run
- Make predictions by sending data to DLHub and specifying the servable to use

```
from dlhub_sdk.client import DLHubClient
dl = DLHubClient()
pred = dl.run("candle_p1b1", data)
```

X-Ray Science

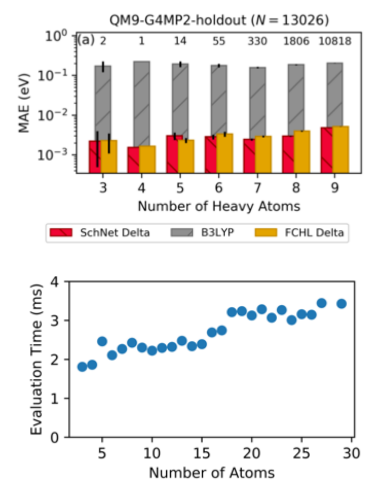
```
from dlhub_sdk.client import DLHubClient
dl = DLHubClient()

struct = dl.run("cherukara_structure", X)
```

CDI Intensity | Predicted Structure (via DLHub)



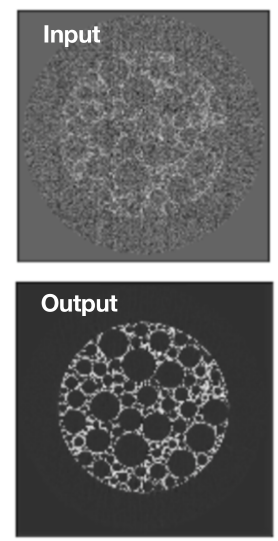
Cherukara et al.

Energy Storage



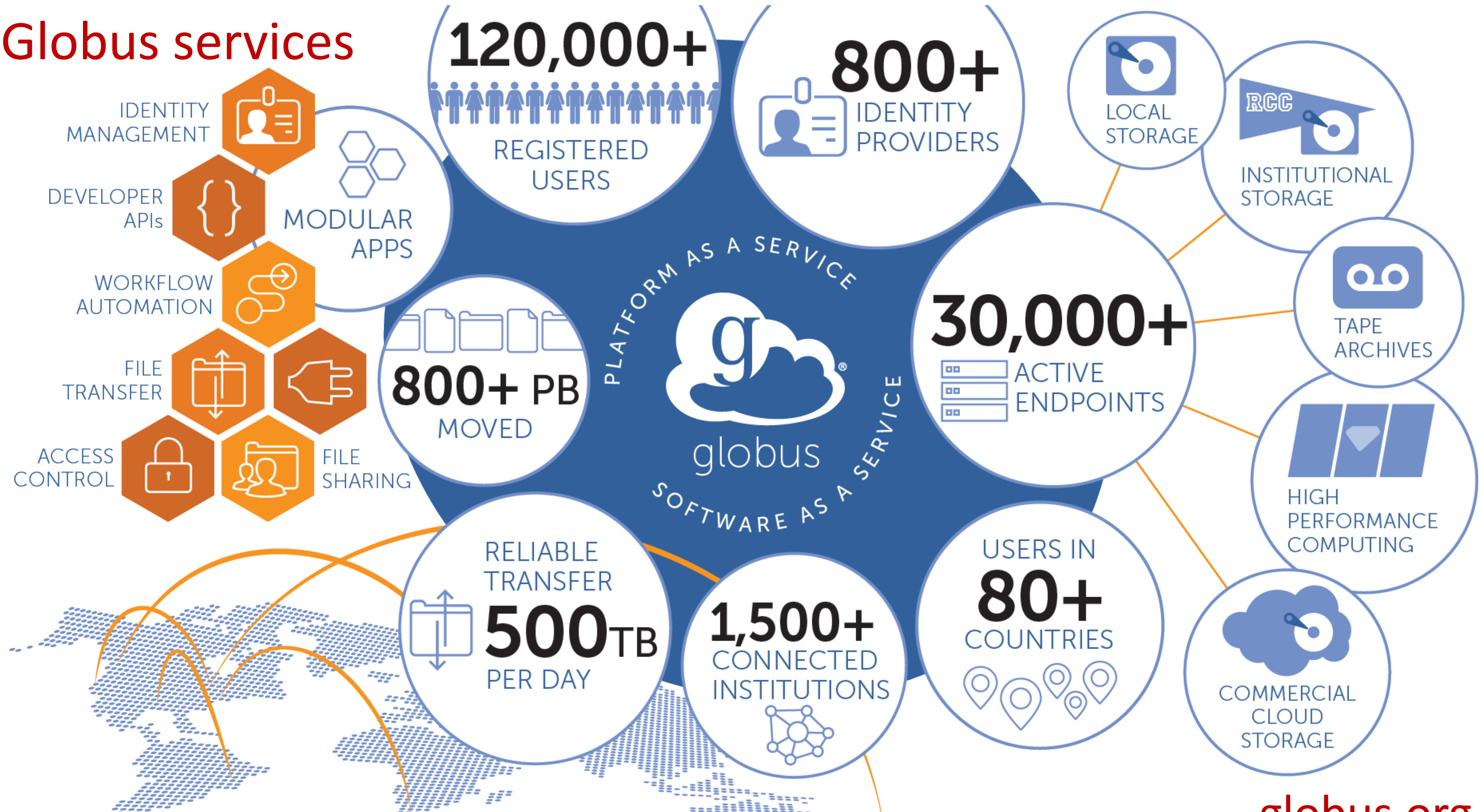
Ward et al.

Tomography



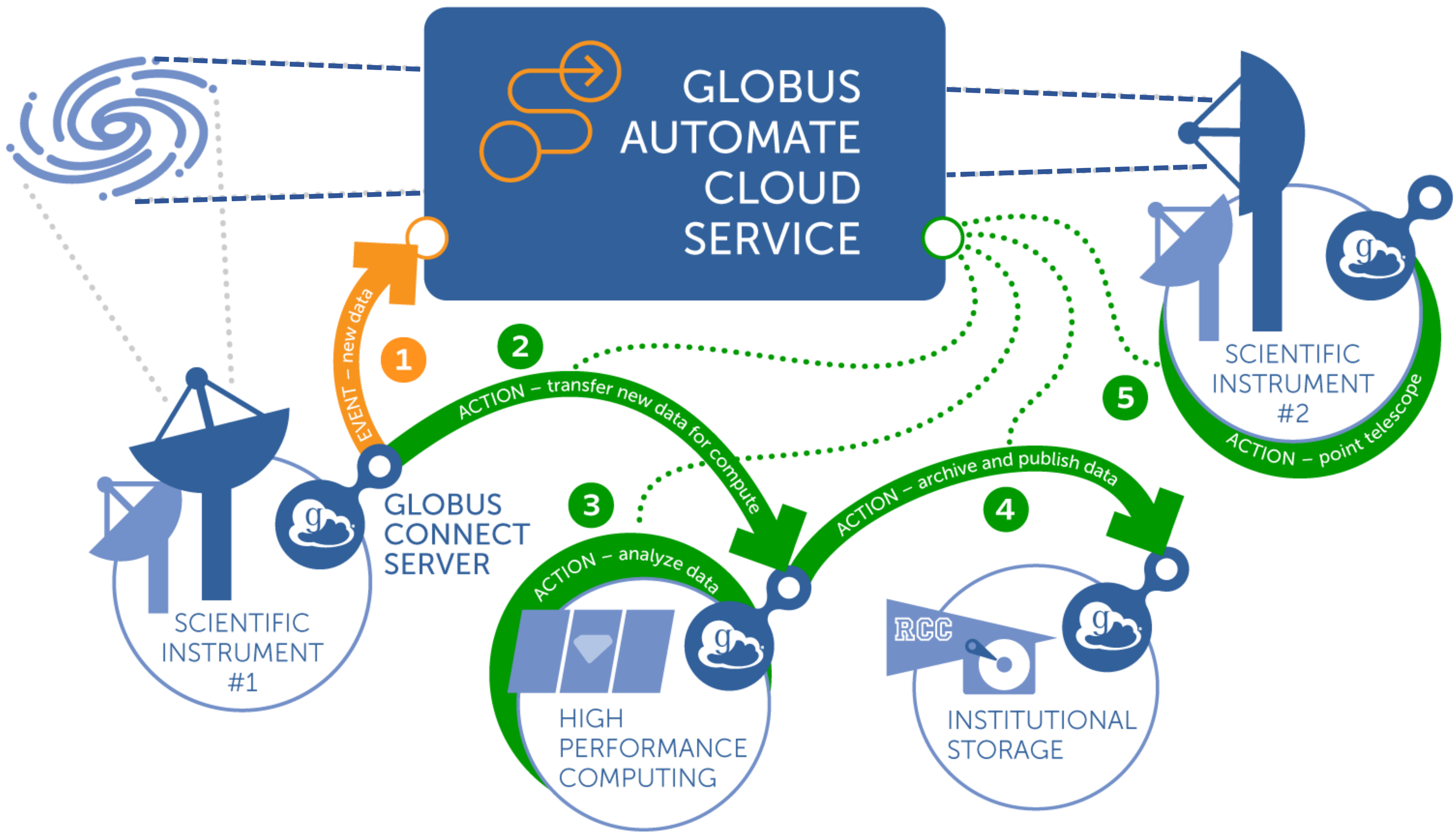
TomoGAN: Liu et al.

Globus services

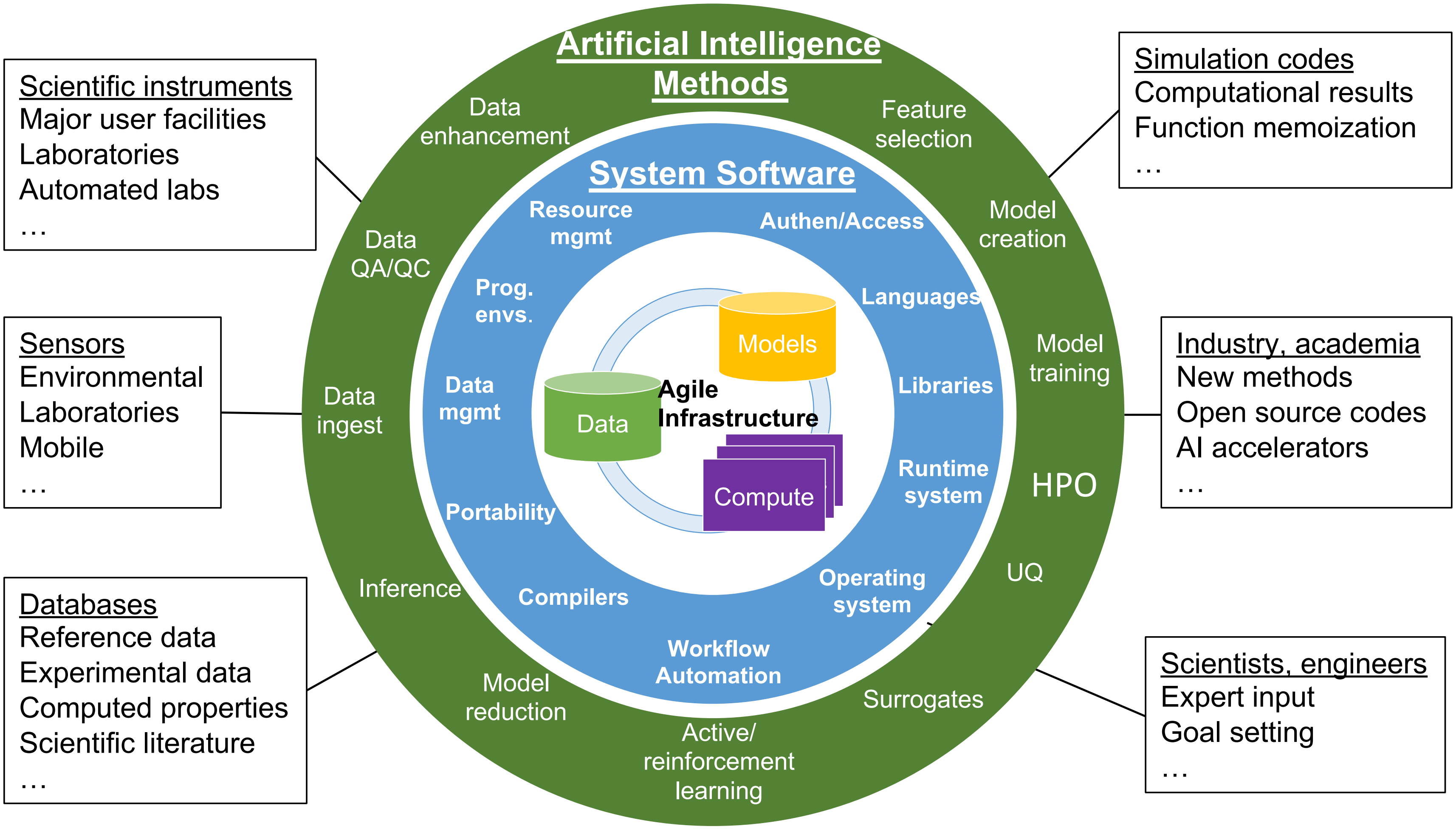


operated by UChicago for researchers worldwide

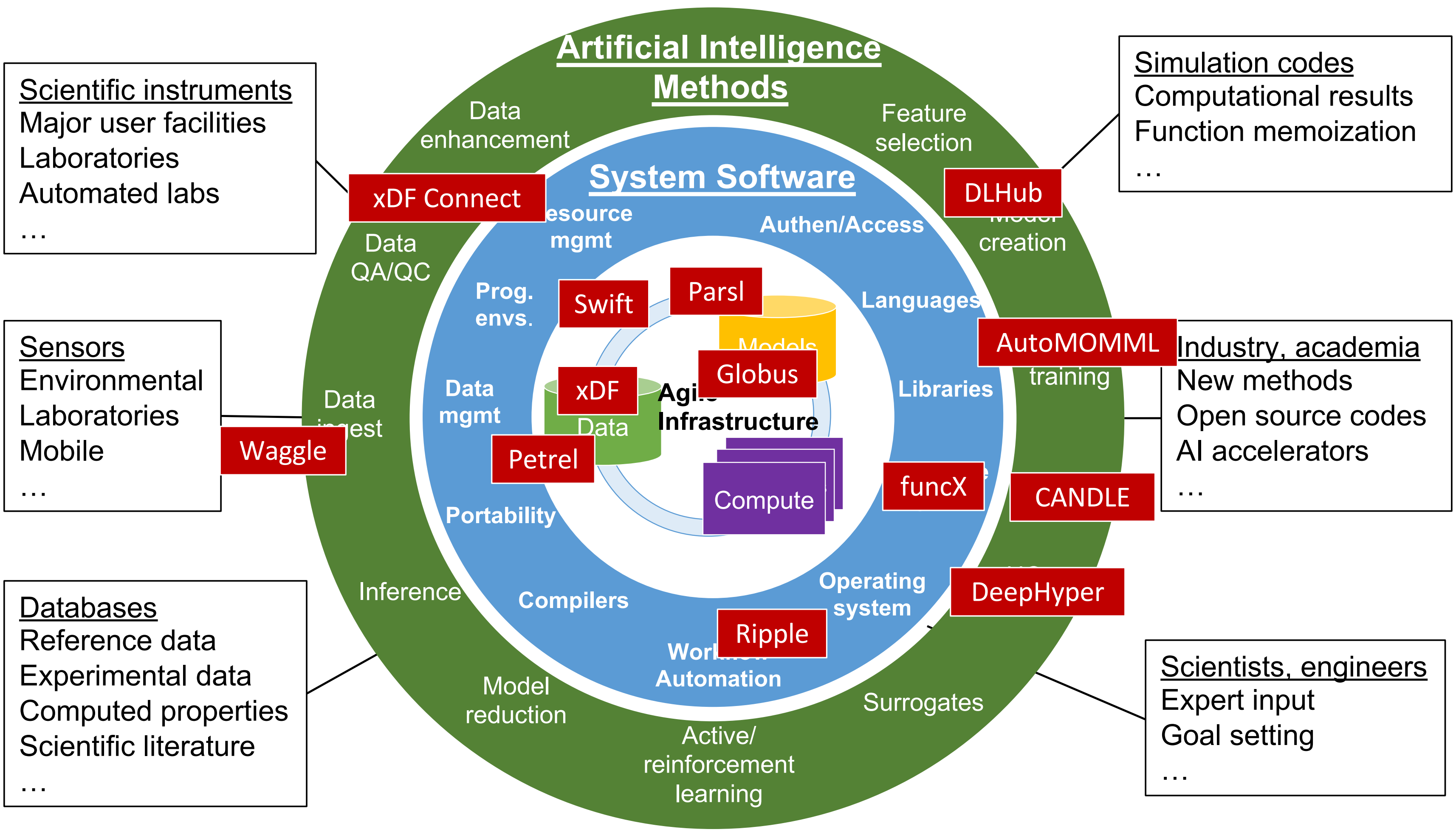
globus.org


GLOBUS
AUTOMATE
CLOUD
SERVICE
EVENT – new data
1
2
ACTION – transfer new data for compute
GLOBUS
CONNECT
SERVER
SCIENTIFIC
INSTRUMENT
#1
3
ACTION – analyze data
HIGH
PERFORMANCE
COMPUTING
ACTION – archive and publish data
4
RCC
INSTITUTIONAL
STORAGE
5
SCIENTIFIC
INSTRUMENT
#2
ACTION – point telescope

We must also rethink other technologies



We must also rethink other technologies



AI for Science @ Argonne targets both research and infrastructure

Applications	AI applications across science and engineering. Transformative approaches to simulation and experimental science.
Learning systems	AI software. Software infrastructure for managing data, models, workflows etc., and for delivering AI capabilities to 1,000s of scientists and engineers.
Foundations	Mathematics, algorithms; general AI, reinforcement learning, uncertainty quantification, explainability, etc.
Hardware	Advanced hardware to support AI. Evaluation of new architectures and systems; exploration of neuromorphic and quantum as long term accelerators for AI.

Future directions in foundations

- Leverage expertise in automatic differentiation, symbolic computing, and optimization to ensure that machine learning for science is forward looking, methods are robust and models interpretable
- Many facets relevant to science
  – Integration of symbolic computing with machine learning
  – Prediction and inference of spatio-temporal processes
  – Derivatives for training, sensitivity analysis, optimization, and UQ
  – Rapid data analysis to reduce volume or identify features of interest
  – Variety of new approaches to inference and UQ
  – Identify and account for uncertainty in data sources and computations

In symbols one observes an advantage in discovery which is greatest when they express the exact nature of a thing briefly and, as it were, picture it; then indeed the labor of thought is wonderfully diminished.

— Gottfried Wilhelm Leibniz

Argonne was the home to a leading symbolic AI group from the 1960s to the mid 2000s working on Automated Theorem Proving

Alan Bundy	University of Edinburgh
Edmund Clarke	Carnegie Mellon University
Tammi Henry	University of Tennessee
Larry Hines	University of Texas
Deepak Kapur	State University of New York at Albany
Matt Kaufmann	Computational Logic
Ken Kunen	University of Wisconsin
Vladimir Lifschitz	Stanford University
Ewing Lusk	Argonne
William McCune	Argonne
Ross Overbeek	Argonne
Dana Scott	Carnegie Mellon University
Mark Stickel	SRI
Rick Stevens	Argonne
Robert Veroff	University of New Mexico
Richard Waldinger	SRI
Steve Winker	Argonne
Larry Wos	Argonne
Hantao Zhang	The University of Iowa

Attendees at an Argonne ATP "theory institute" in 1990.

1. Algebraic Geometry
    - Cancellative Semigroups on a Cubic Curve
    - Uniqueness of the 5-ary Steiner Law
2. Cancellative Semigroups
3. Lattice Theory
    - A Simpler Absorptive Basis for Lattice Theory
    - A New Schema for Single Axioms
    - A Shorter Single Axiom for Lattice Theory
    - A Single Axiom for Weakly Associative Lattices
4. Quasilattice Theory
5. Uniqueness of Operations in Lattice-like Algebras
6. Self-dual Bases for Boolean Algebra
7. Self-dual 2-Basis for Group Theory
8. Self-dual Bases for Group Varieties
9. Quasigroup Theory
10. Quasigroup Design Problems
11. Single Axioms for Ternary Boolean Algebra
12. Single Axioms for Groups
    - Ordinary Groups
    - Abelian Groups
    - Exponent Groups
    - Some Permutative Varieties
    - Ordinary Groups (Kunen)
    - Groups of Exponent 4 (Kunen)
    - Odd Exponent Groups (Hart and Kunen)
13. Simple Bases for Moufang Loops
14. Single Axioms for Inverse Loops and Subvarieties
15. Left Group and Right Group Calculi
16. Fixed Point Combinators
17. Semigroup Structure (F3B2)
18. Illative Combinatory Logic (Jech)
19. Robbins Algebra and Boolean Algebra
20. Equivalential Calculus Single Axioms
21. Semigroups, Antiautomorphisms, and Involutions
22. Independence of Ternary Boolean Algebra Axioms
23. Two-valued Sentential Calculus
24. Many-valued Sentential Calculus
25. Short Proofs in Various Logic Calculi
26. Pure Proofs in Equivalential Calculus

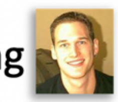
Credit: Lloyd DeGrane / The New York Times

Dr. William McCune at Argonne Labs, Illinois in his office with computer. The "Proof of Robbins Conjecture" problem is on the screen.

1996

8 Days on an IBM RS/6000
30 megabytes of memory

Powerful satisfiability modulo theory (SMT) solvers like Microsoft's Z3 adapt strategies from Otter

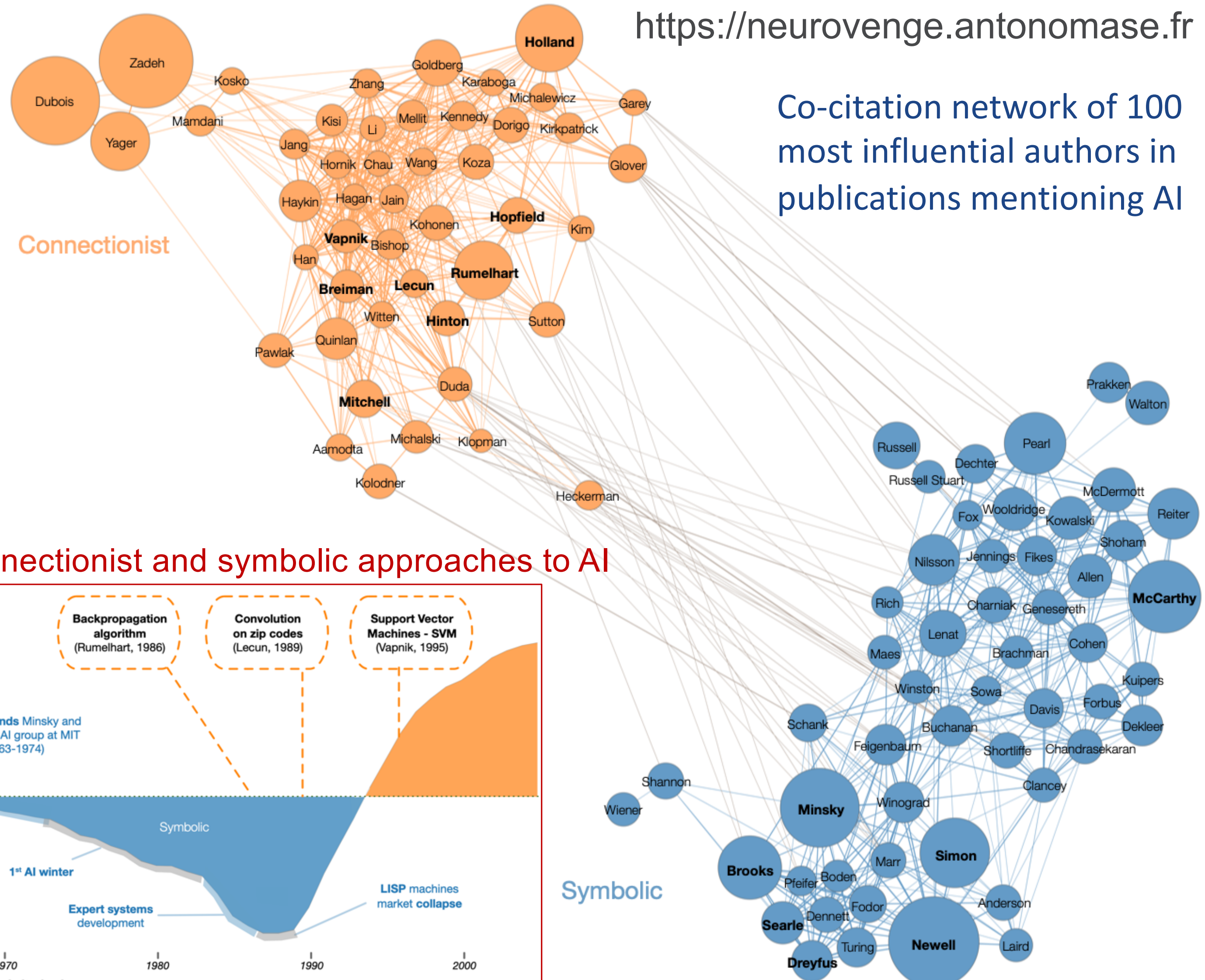
- Bounded model-checking of model programs 
- Termination 
- Security protocols  
- Business application modeling 
- Cryptography 
- Model Based Testing (SQL-Server)
- *Your killer-application here*

Sources: https://nyti.ms/37X8fUM, http://bit.ly/35J3lcE

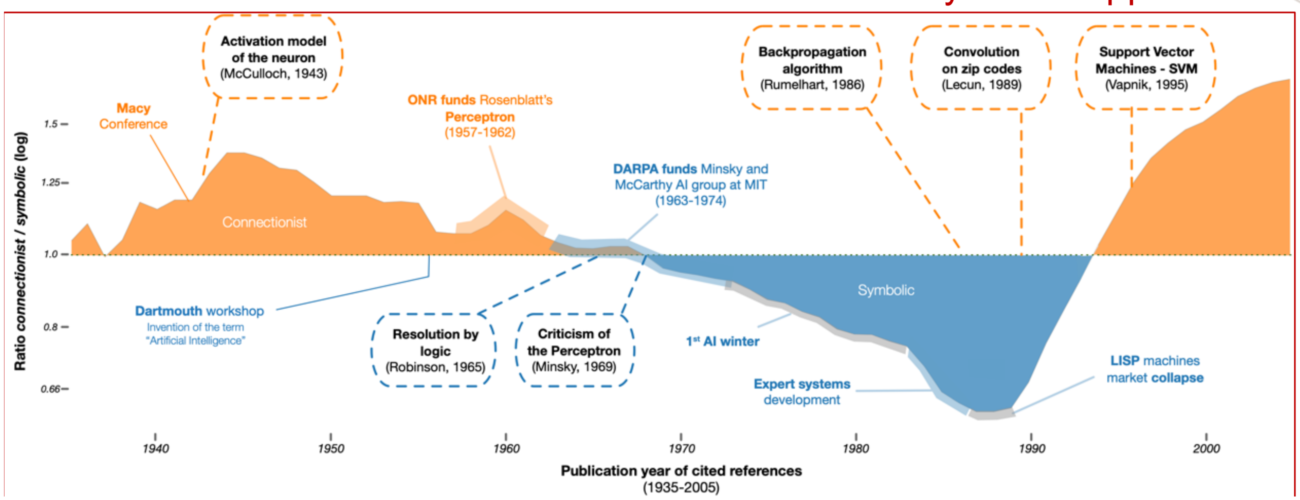
AI research has strong symbolic as well as connectionist roots

https://neurovenge.antonomase.fr

Co-citation network of 100 most influential authors in publications mentioning AI



Evolution of academic influence of connectionist and symbolic approaches to AI

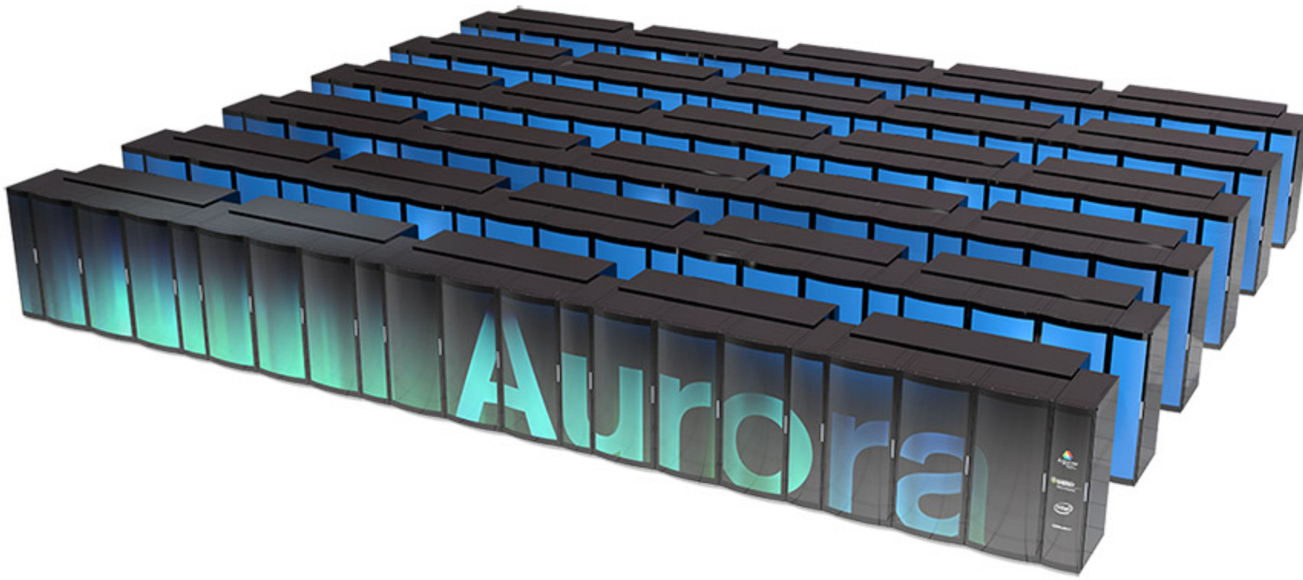


AI for Science @ Argonne targets both research and infrastructure

Applications	AI applications across science and engineering. Transformative approaches to simulation and experimental science.
Learning systems	AI software. Software infrastructure for managing data, models, workflows etc., and for delivering AI capabilities to 1,000s of scientists and engineers.
Foundations	Mathematics, algorithms; general AI, reinforcement learning, uncertainty quantification, explainability, etc.
Hardware	Advanced hardware to support AI. Evaluation of new architectures and systems; exploration of neuromorphic and quantum as long term accelerators for AI.

Aurora: HPC and AI

>> Exaops/s for AI



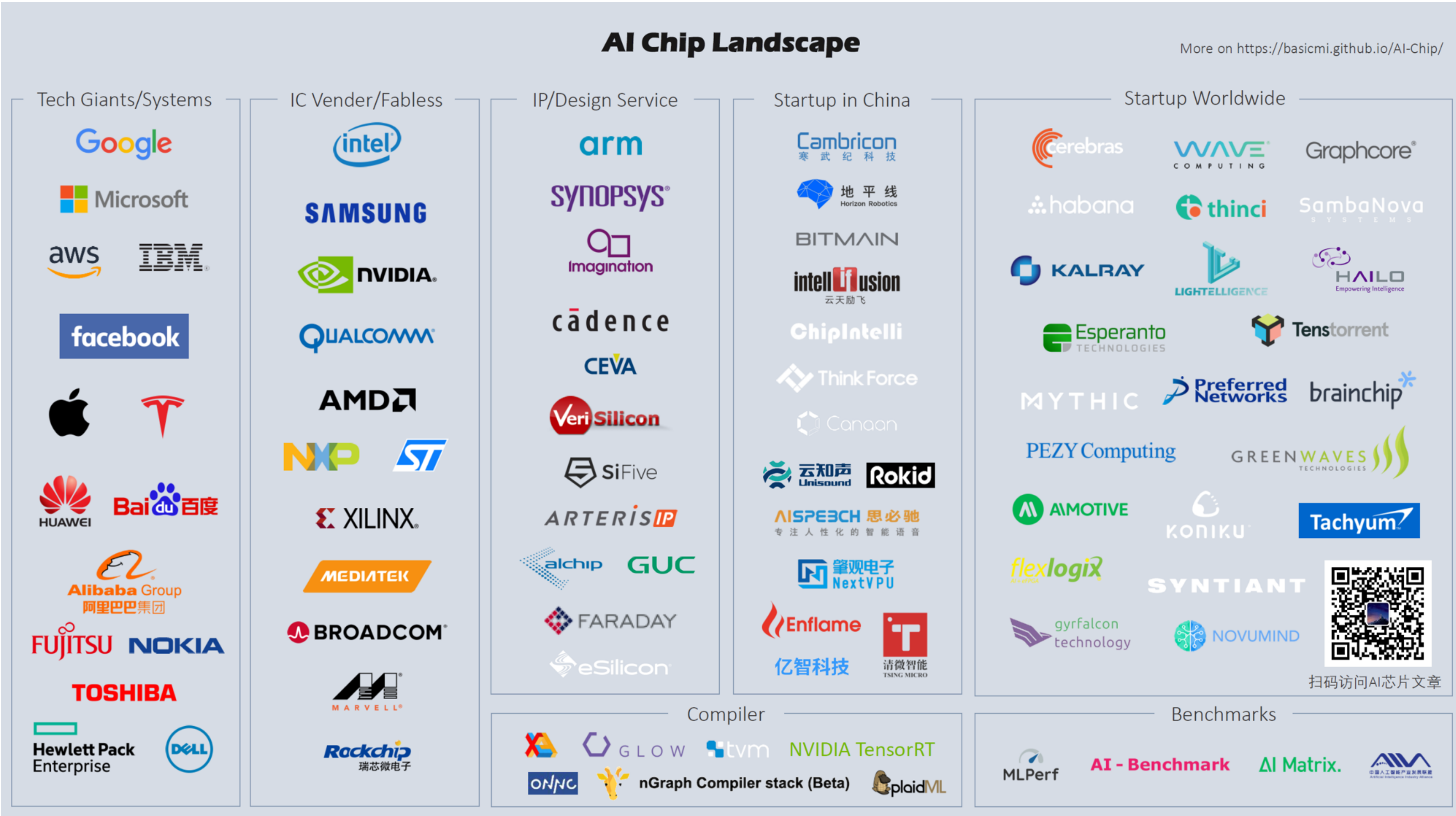
Architecture supports three types of computing

- 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

VCs investing >$4B in startups for AI acceleration

Which platforms will be good for science?


AI Chip Landscape
More on https://basicmi.github.io/AI-Chip/
Tech Giants/Systems
Google
Microsoft
aws
IBM
facebook
HUAWEI
Baidu 百度
Alibaba Group
阿里巴巴集团
FUJITSU
NOKIA
TOSHIBA
Hewlett Packard Enterprise
DELL
IC Vender/Fabless
intel
SAMSUNG
NVIDIA
QUALCOMM
AMD
NXP
ST
XILINX
MEDIATEK
BROADCOM
MARVELL
Rockchip
瑞芯微电子
IP/Design Service
arm
synopsys
Imagination
cadence
CEVA
VeriSilicon
SiFive
ARTERIS IP
alchip
GUC
FARADAY
eSilicon
Startup in China
Cambricon
寒 武 纪 科 技
地 平 线
Horizon Robotics
BITMAIN
intellifusion
云天励飞
ChipIntelli
Think Force
Canaan
云知声
Unisound
Rokid
AISPEECH 思必驰
专 注 人 性 化 的 智 能 语 音
肇观电子
NextVPU
Enflame
亿智科技
清微智能
TSING MICRO
Startup Worldwide
cerebras
WAVE
COMPUTING
Graphcore
habana
thinci
SambaNova
SYSTEMS
KALRAY
LIGHTELLIGENCE
HAILO
Empowering Intelligence
Esperanto
TECHNOLOGIES
Tenstorrent
MYTHIC
Preferred Networks
brainchip
PEZY Computing
GREENWAVES
TECHNOLOGIES
AIMOTIVE
KONIKU
Tachyum
flexlogix
SYNTIANT
gyrfalcon technology
NOVUMIND
扫码访问AI芯片文章
Compiler
GLOW
tvm
NVIDIA TensorRT
ONNC
nGraph Compiler stack (Beta)
plaidML
Benchmarks
MLPerf
AI - Benchmark
AI Matrix.

An AI accelerator testbed

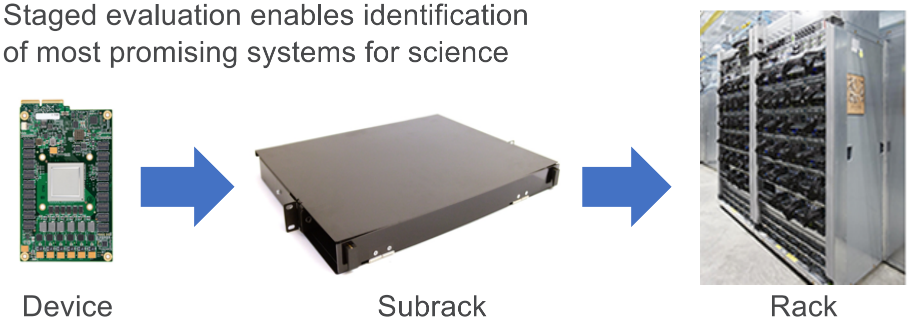
Engaging the community to understand and improve specialized AI hardware for science

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

IC Vendors	Intel, Qualcomm, Nvidia, Samsung, AMD, Xilinx, IBM, STMicroelectronics, NXP, Marvell, MediaTek, HiSilicon, Rockchip	13
Tech Giants & HPC Vendors	Google, Amazon_AWS, Microsoft, Apple, Aliyun, Alibaba Group, Tencent Cloud, Baidu, Baidu Cloud, HUAWEI Cloud, Fujitsu, Nokia, Facebook, HPE, Tesla	12
IP Vendors	ARM, Synopsys, Imagination, CEVA, Cadence, VeriSilicon, Videantis	7
Startups in China	Cambricon, Horizon Robotics, Bitmain, Chipintelli, Thinkforce, Unisound, AISpeech, Rokid, NextVPU, Canaan, Enflame, Eesay Tech	12
Startups Worldwide	Cerebras, Wave Computing, Graphcore, PEZY, Tenstorrent, ThinCI, Koniku, Adapteva, Knowm, Mythic, Kalray, BrainChip, Almotive, DeepScale, Leepmind, Krtkl, NovuMind, REM, TERADEEP, DEEP VISION, Groq, KAIST DNPU, Kneron, Esperanto Technologies, Gyrfalcon Technology, SambaNova Systems, GreenWaves Technology, Lightelligence, Lightmatter, ThinkSilicon, Innogrit, Kortiq, Hailo,Tachyum,AlphaICs,Syntiant, Habana, aiCTX, Flex Logix, Preferred Network, Cornami, Anaflash, Optaylsys, Eta Compute	44

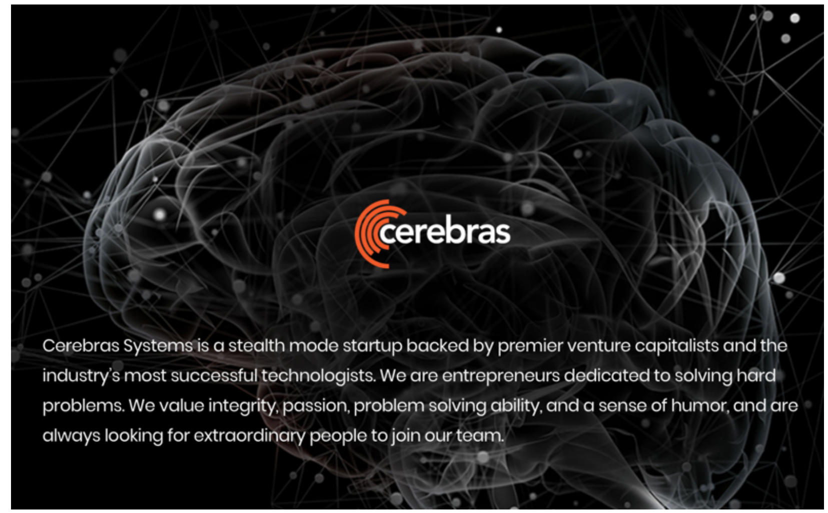
Staged evaluation enables identification of most promising systems for science



https://github.com/basicmi/AI-Chip

Argonne is developing AI infrastructure

- Argonne is partnering with Cerebras to develop and deploy an AI computing platform
- Scientific AI models from cancer, cosmology, brain imaging, and materials science are the first examples that will be deployed
- Our goal is to accelerate relevant AI model types for problems in materials, biomedical, cosmology, high-energy physics, energy systems, synthetic biology, climate, software optimization, architecture research etc.



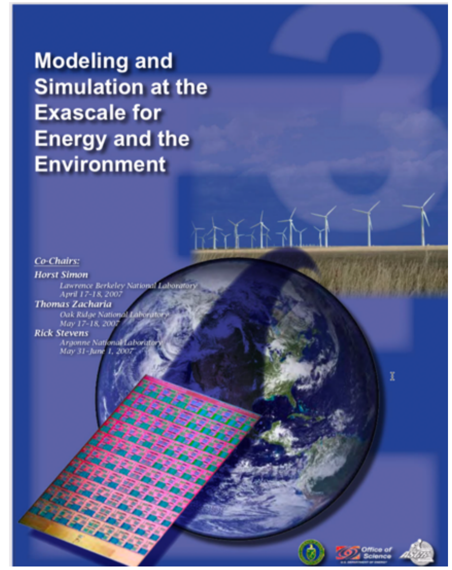
CONCLUSION: FUTURE cerebras

- Massive multi-core engines that **enable model parallelism**
- Orders of magnitude **greater memory and communication BW**
- Unconstrained methods, e.g., large and **small mini-batch**
- Capture weight and activation **sparsity** for higher performance
- Support research and execution of **emergent model architectures** (not just those of today)

DOE's AI for Science Townhalls

Organized by Argonne, Oak Ridge and Berkeley with participation from all DOE labs

- 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.
- ~300 people per meeting
- Output will be a report to guide strategic planning at Labs and DOE



I have covered just a few of the many activities underway at Argonne on AI in science

We look forward to collaboration with Fermilab and Uchicago in applications, learning systems, foundations, and hardware

Ian Foster
foster@anl.gov

A 20-Year Community Roadmap for Artificial Intelligence Research in the US

http://bit.ly/2JK8OZ9

CCC
Computing Community Consortium
Catalyst

AAAI
Association for the Advancement of Artificial Intelligence