AI for Science

Dr. Alok Choudhary
Henry and Isabel Dever Professor
EECS and Kellogg School of Management
Northwestern University
choudhar@eecs.northwestern.edu

Founder, Chairman and Chief Scientist
4Cinsights Inc: A Big Data Science Company
(recently acquired by Mediaocean)
www.4cinsights.com

Secretary of Energy Advisory Board AI Working Group member

Keynote @ IEEE Cluster2020 September 14-17, 2020, Kobe, Japan

http://cucis.ece.northwestern.edu/publications/

Discovery and Design Paradigms

1st paradigm: ***Empirical science***

Experiments

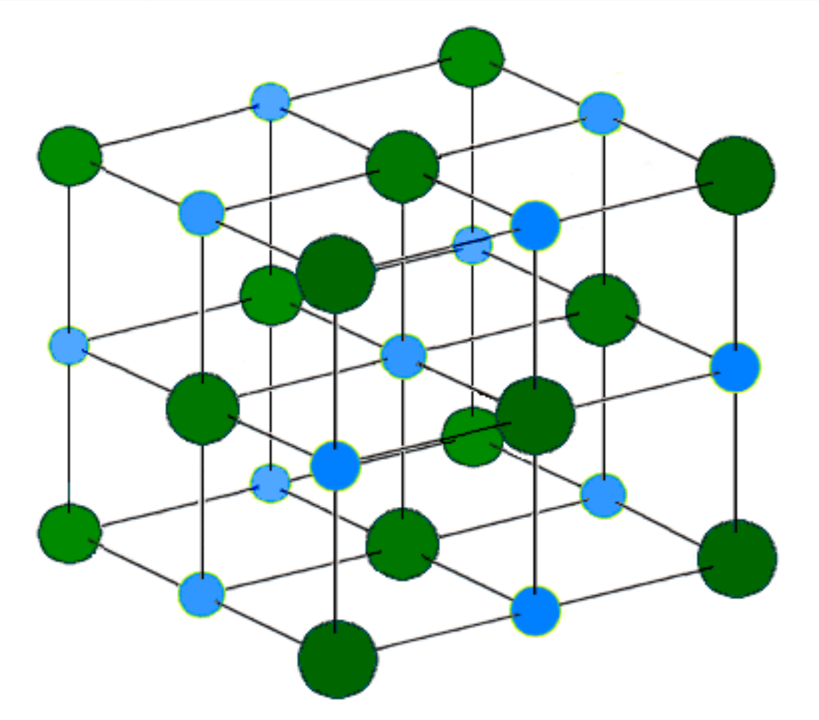
2nd paradigm: ***Model-based theoretical science***

$$\Delta U = Q - W$$

Change in internal energy | Heat added to system | Work done by system

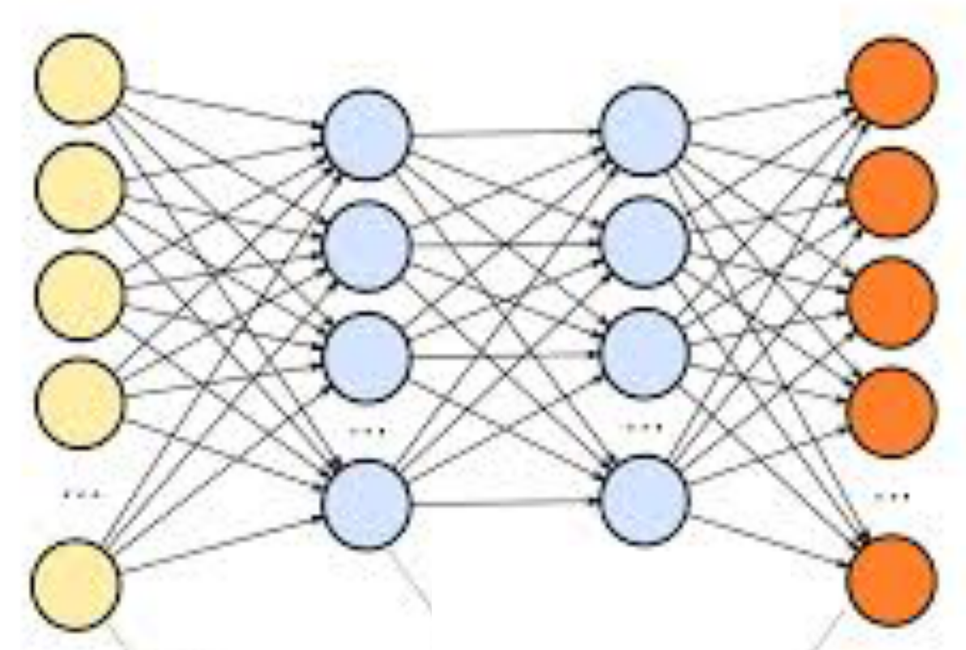
Laws of Thermodynamics

3rd paradigm: ***Computational science (simulations)***



Density Functional Theory, Molecular Dynamics

4th paradigm: ***(Big) data driven science***



Predictive analytics
Clustering
Relationship mining
Anomaly detection

?

1600 | 1950 | 2000

Big Data + Big Compute = An Extreme Scale Marriage for Smarter Science?

Alok Choudhary
Henry and Isabel Dever Professor
EECS and Kellogg School of Management
Northwestern University

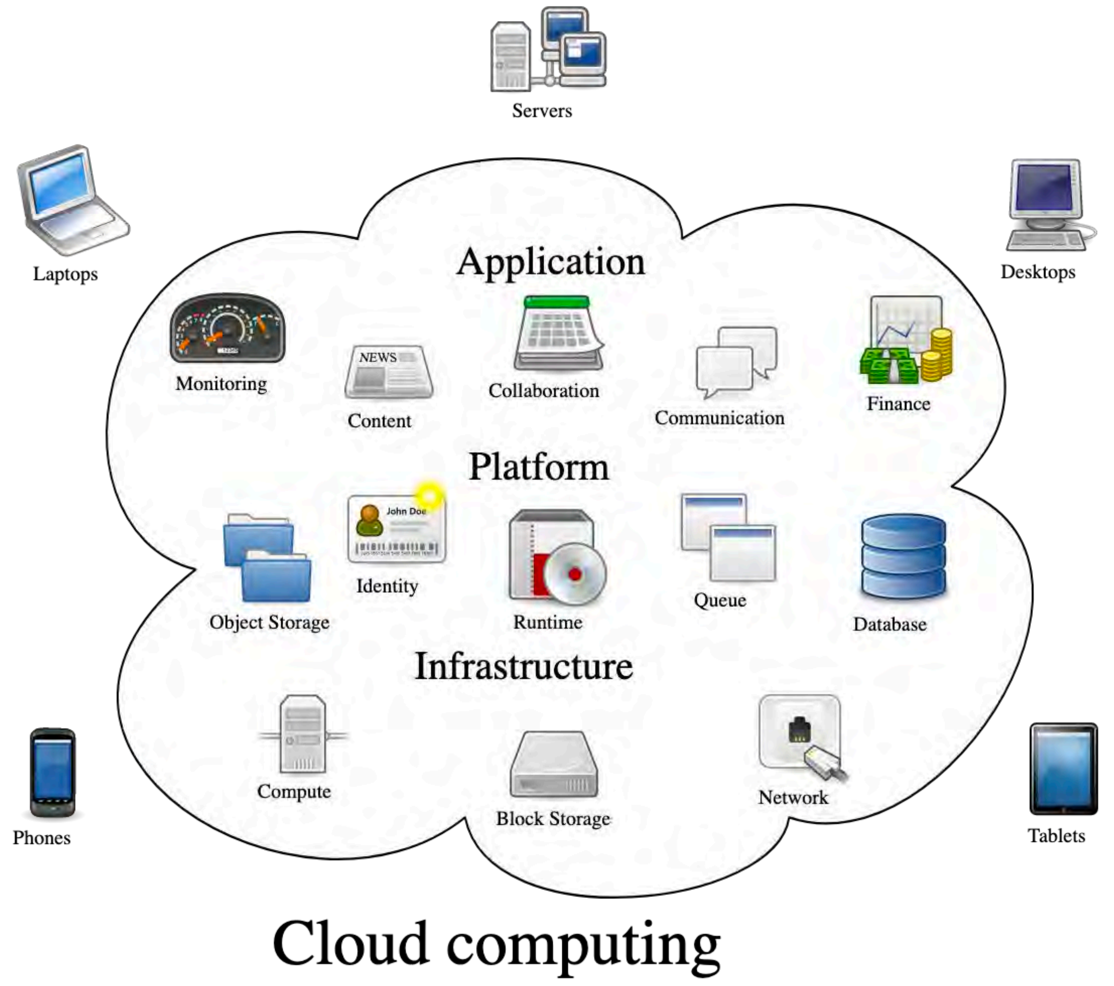
Plenary, SC 2013 , November 21, 2013

Outline

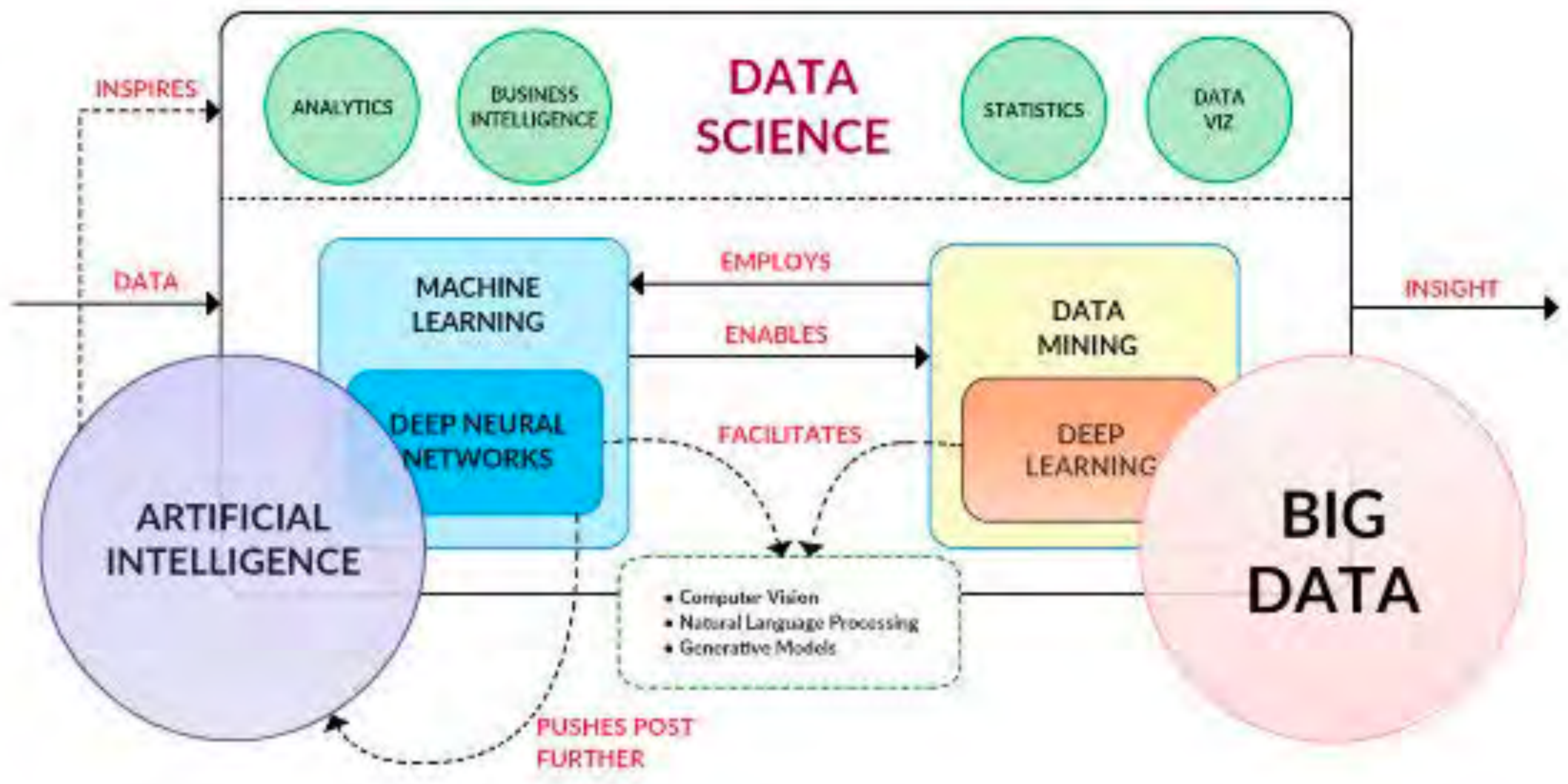
- Why Now?
- AI for Science Premise
- Integration approaches for AI in Discovery Paradigms
- Examples
  - Materials
  - Climate understanding
  - Cosmology
- What's Next?

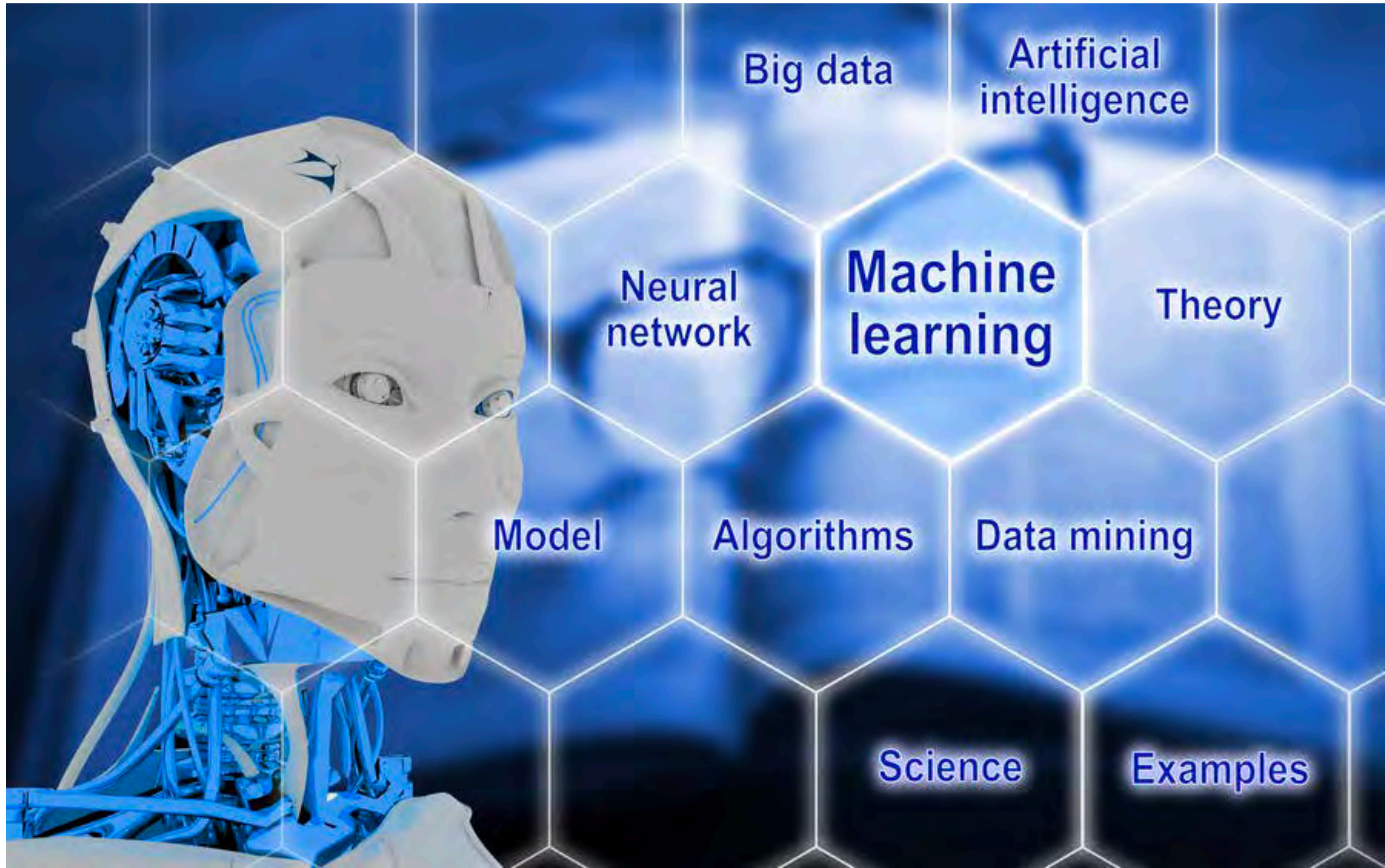
Development 1: HPC + Accelerators

Development 2: Democratization via Cloud Computing

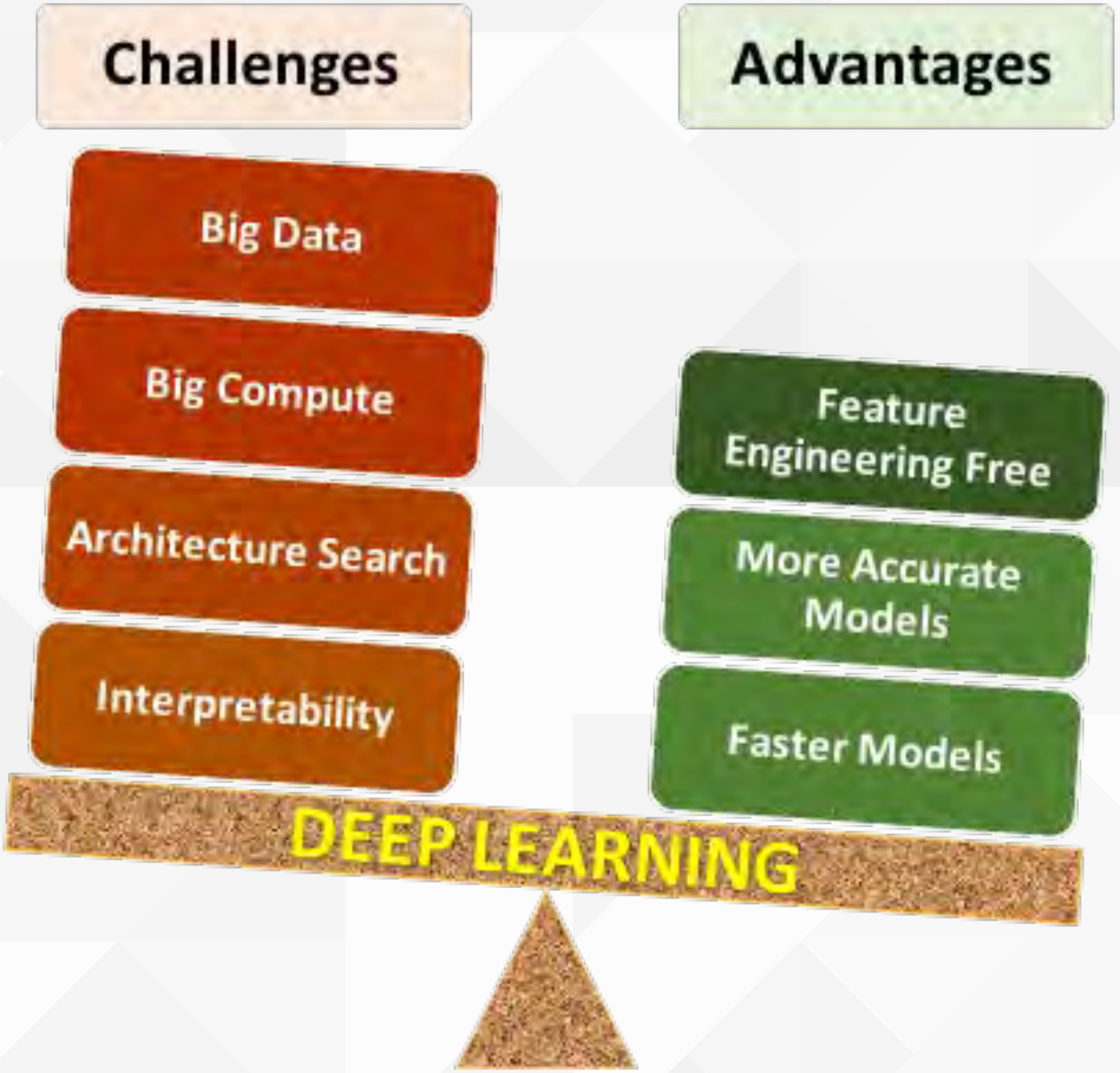


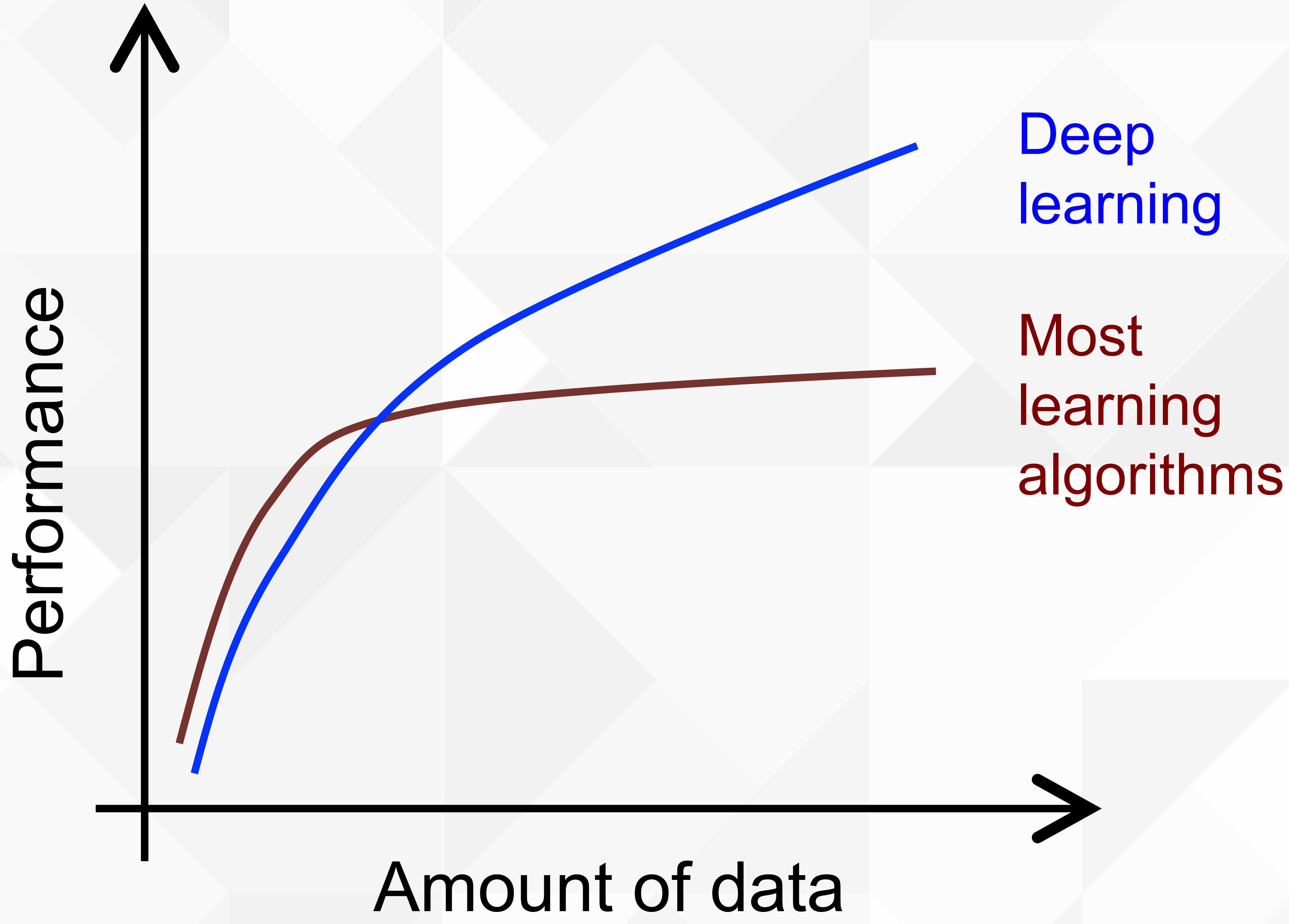
Development 3: ML/AI



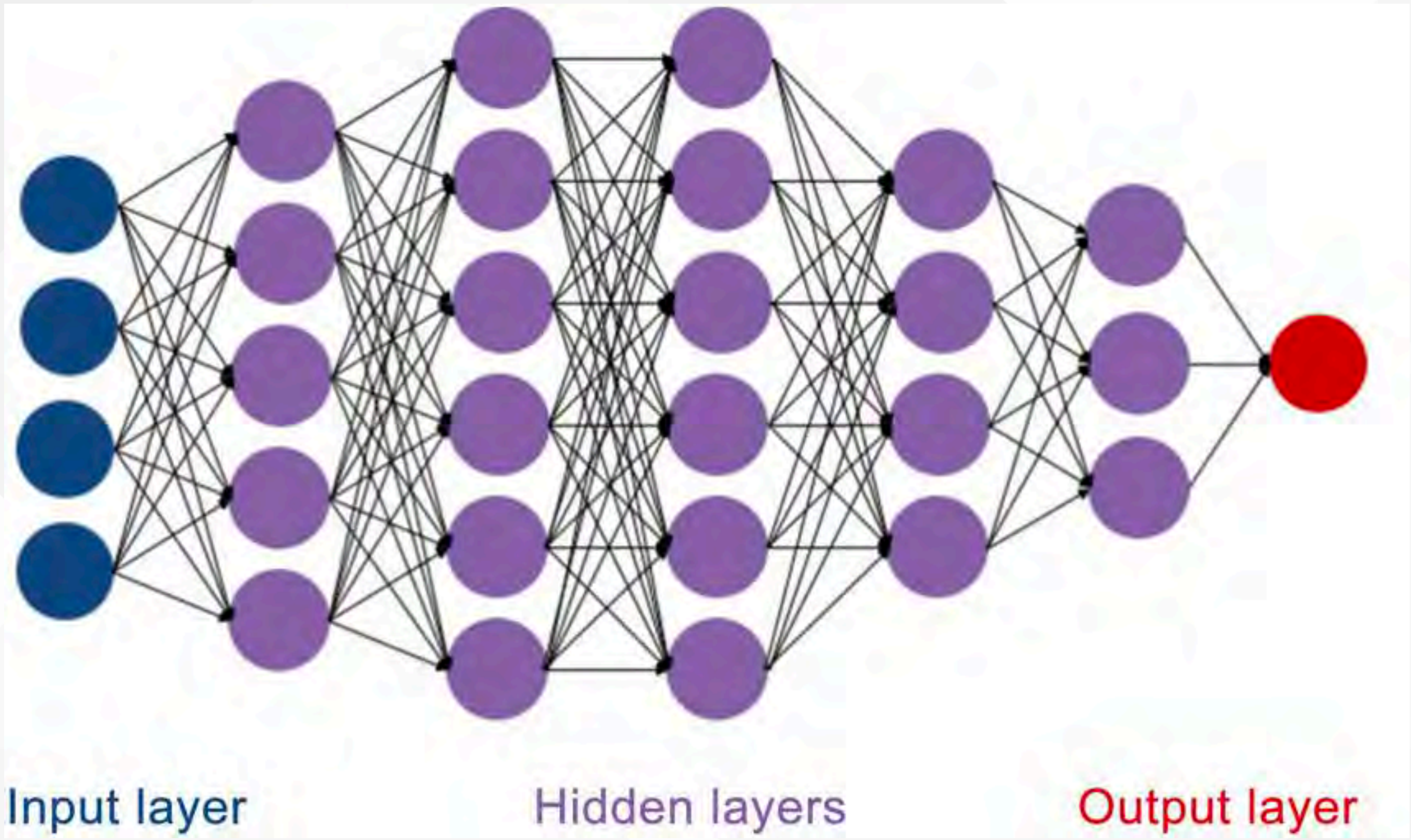


AI Core - Deep Learning

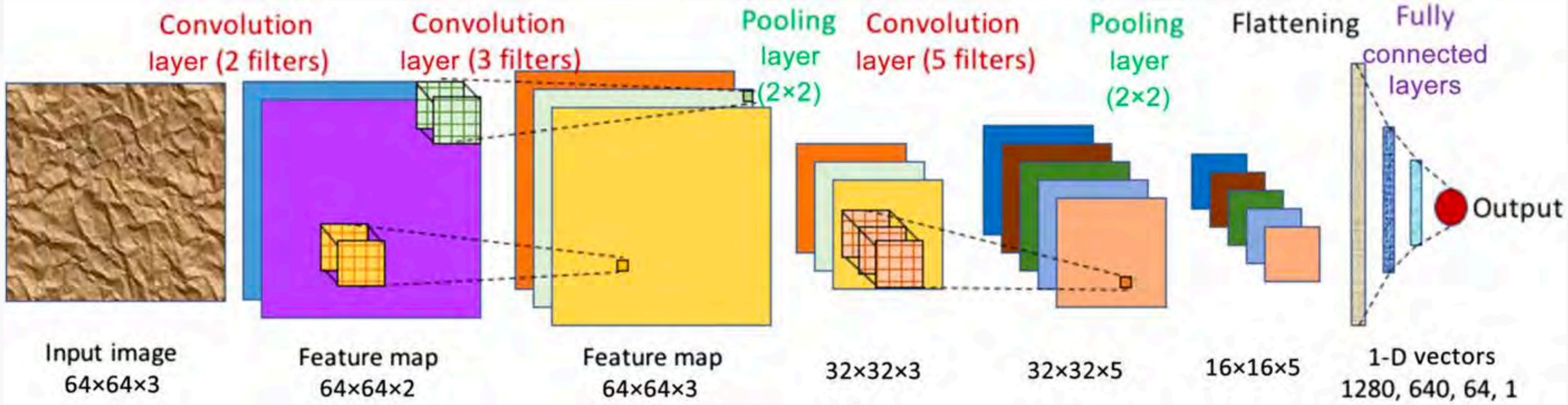




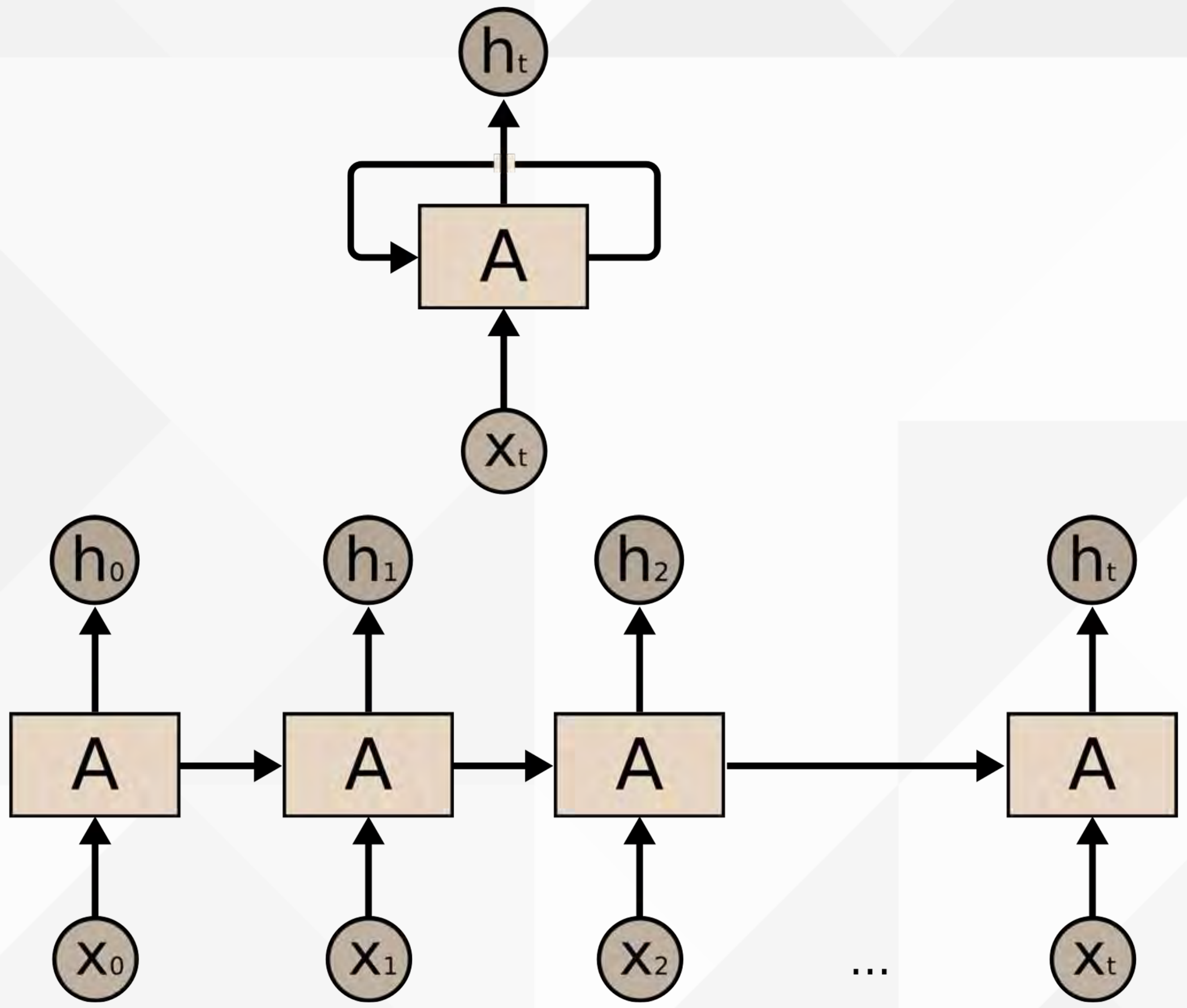
Types of Deep Learning Networks



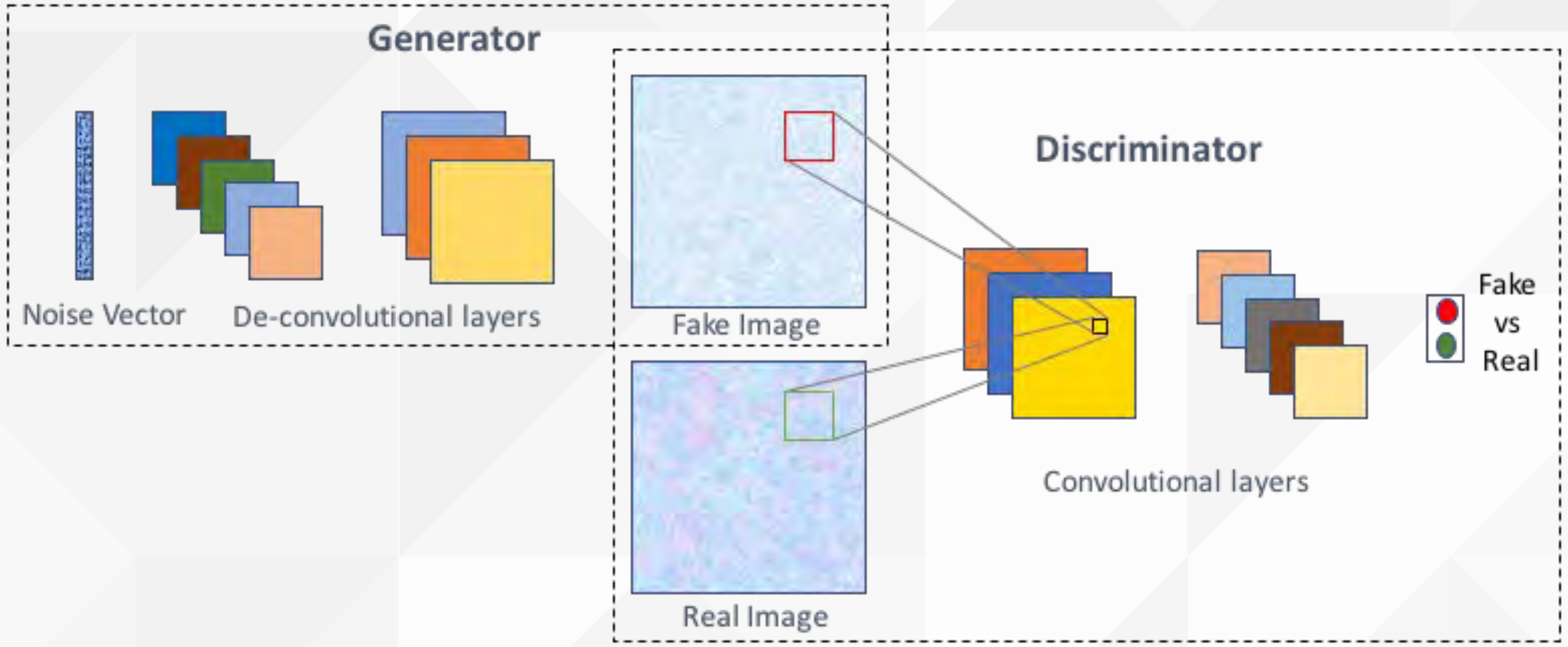
Fully connected network (MLP)



Convolutional neural network (CNN)



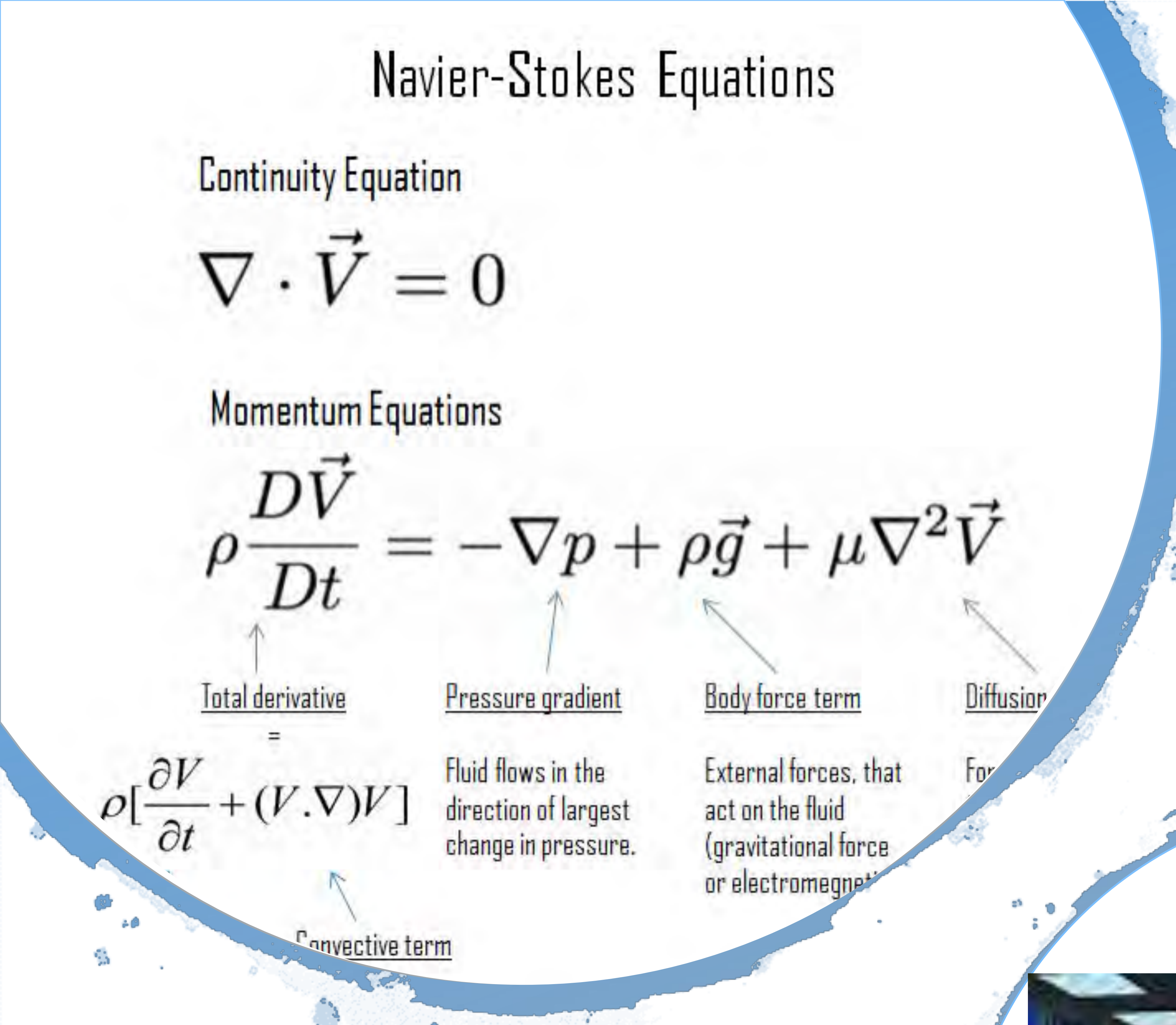
Recurrent neural network (RNN)

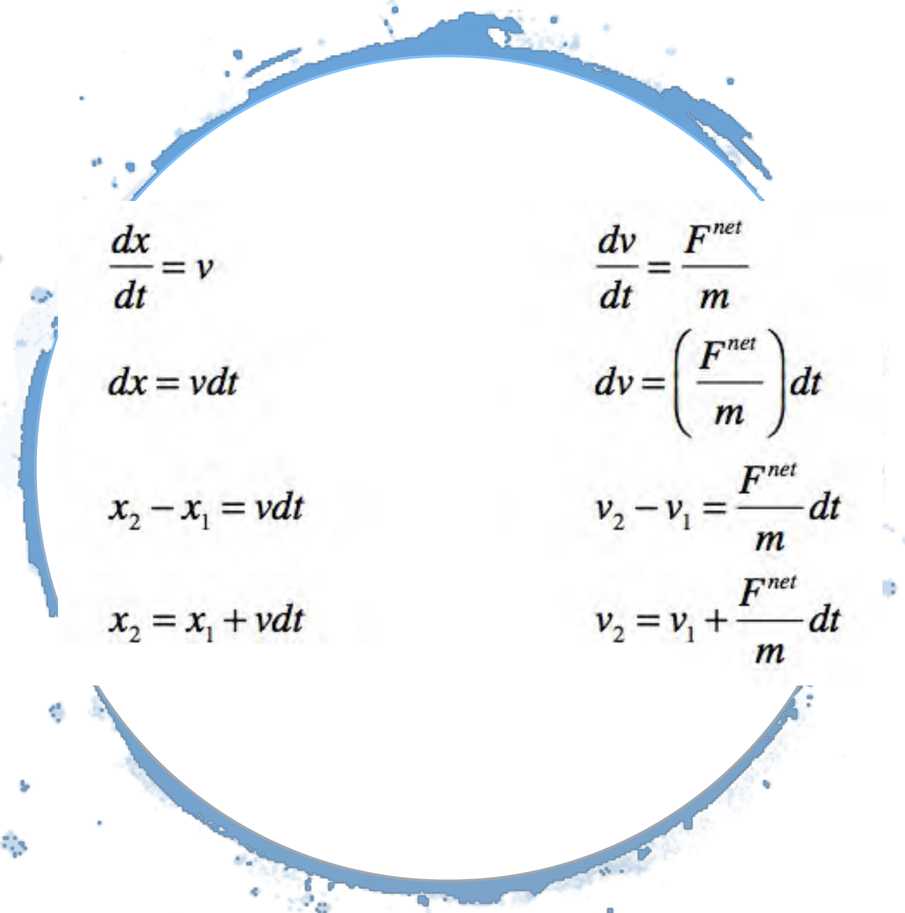


Generative adversarial network (GAN)

AI Premise for Science (and Design)

- Accelerate scientific discoveries by
  - Enabling multiple paradigms to work in concert by accentuating their strengths and overcoming their limitations via Machine Learning
  - A Virtuous Relationship
    - HPC: Enables AI/ML and Big Data Science
    - AI/ML: Accelerates HPC systems designs
    - Cloud: Makes HPC and ML available to everyone
    - HPC+AI: Enables Simulations and Data Science to work in concert
      - accelerating discoveries,
      - prioritizing experiments, designs
  - Complements human strengths



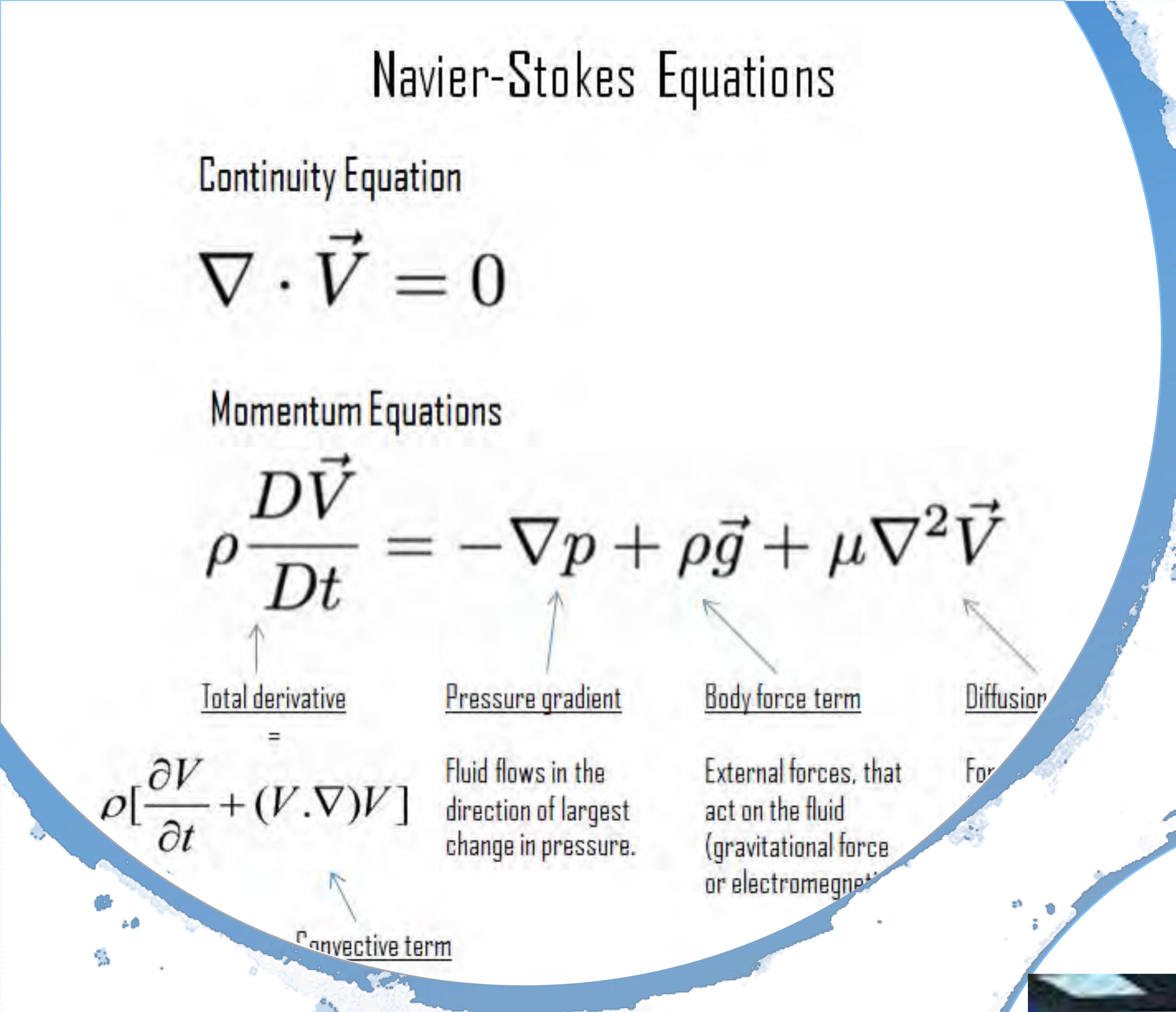


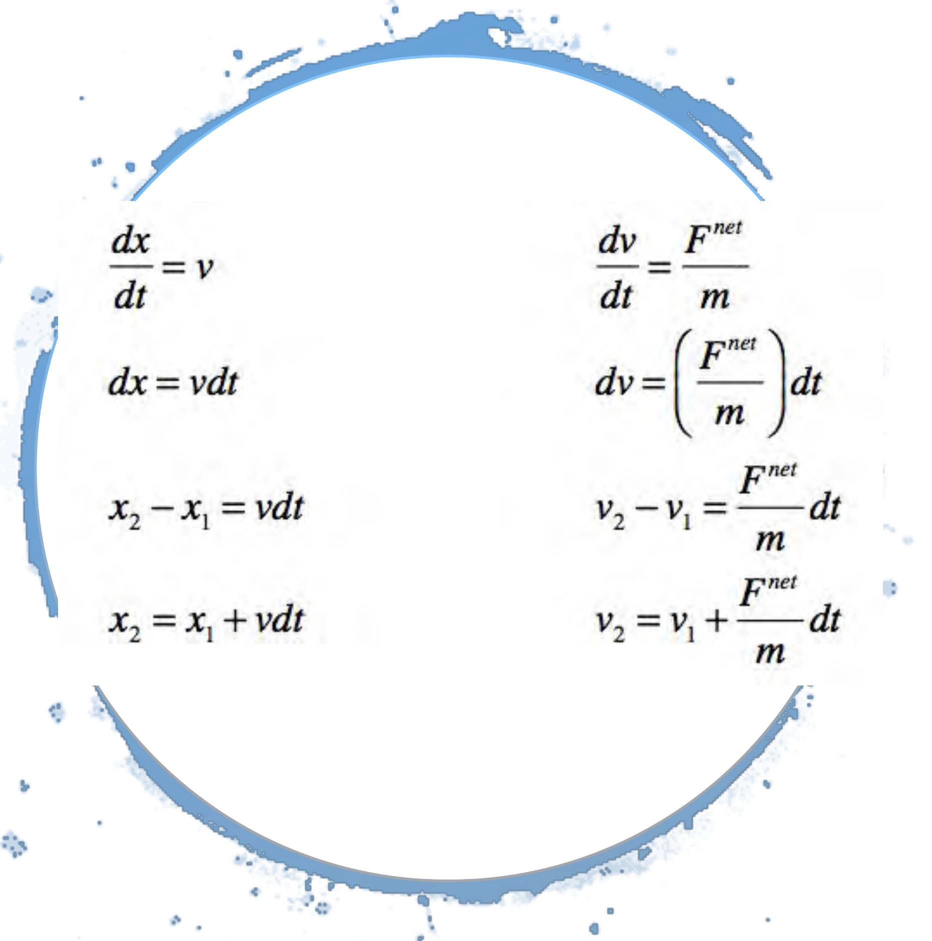
Theory/Model Driven Point Workflow

Theory/Model

Simulations

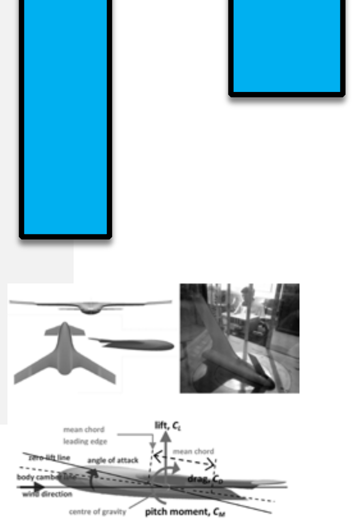
Experiments/Observations





Theory/Model Driven Point Workflow => Point Solution

- Theory/Model
- Simulations
- Experiments/Observations



$$\frac{dx}{dt} = v \qquad \frac{dv}{dt} = \frac{F^{net}}{m}$$

$$dx = vdt \qquad dv = \left(\frac{F^{net}}{m}\right)dt$$

$$x_2 - x_1 = vdt \qquad v_2 - v_1 = \frac{F^{net}}{m}dt$$

$$x_2 = x_1 + vdt \qquad v_2 = v_1 + \frac{F^{net}}{m}dt$$

Experiment/Observation Driven Point Workflow

- Experiments/Observations
- Theory/Model
- Simulations

$$\frac{dx}{dt} = v \qquad \frac{dv}{dt} = \frac{F^{net}}{m}$$

$$dx = vdt \qquad dv = \left(\frac{F^{net}}{m}\right)dt$$

$$x_2 - x_1 = vdt \qquad v_2 - v_1 = \frac{F^{net}}{m}dt$$

$$x_2 = x_1 + vdt \qquad v_2 = v_1 + \frac{F^{net}}{m}dt$$

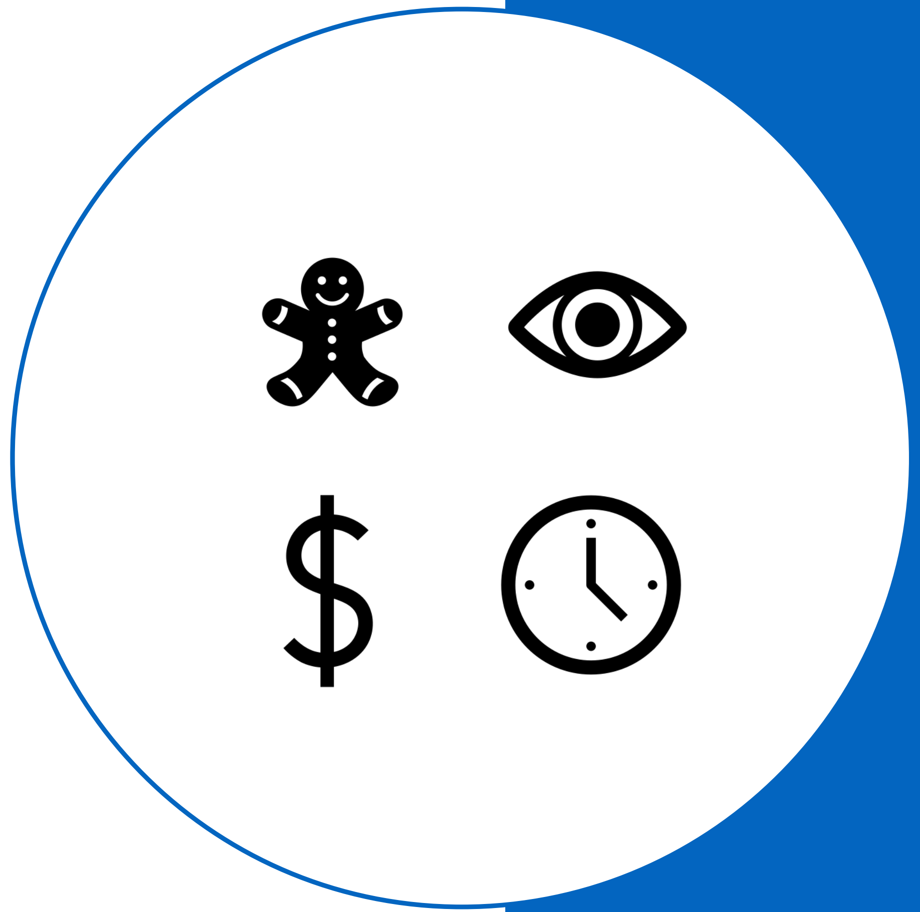
Experiment/Observation Driven Point Workflow

- Experiments/Observations
- Theory/Model
- Simulations

So, what doors does AI open?

What doors does AI open?

- Accelerate discovery of "known unknowns" by leveraging data generated via model-driven point workflows
  - Transforms model-driven science to a predictive modeling science
    - E.g., Discovering properties of materials
- Accelerate discovery of "unknown unknowns" where development of models is difficult, or experiments are infeasible or very expensive
  - Transforms top-down science to a bottom-up discovery process
    - E.g., Inverse models or goal-based designs, learning from data
  - Enables generation of artificial data closely mimicking reality
    - E.g., Cosmology



AI/ML for Science

Multiple Theory/Model

Data from Thousands of Simulations + experiments/observations

AI/ML - Predictive Modeling

Insights

New Experiments or Simulations

Examples

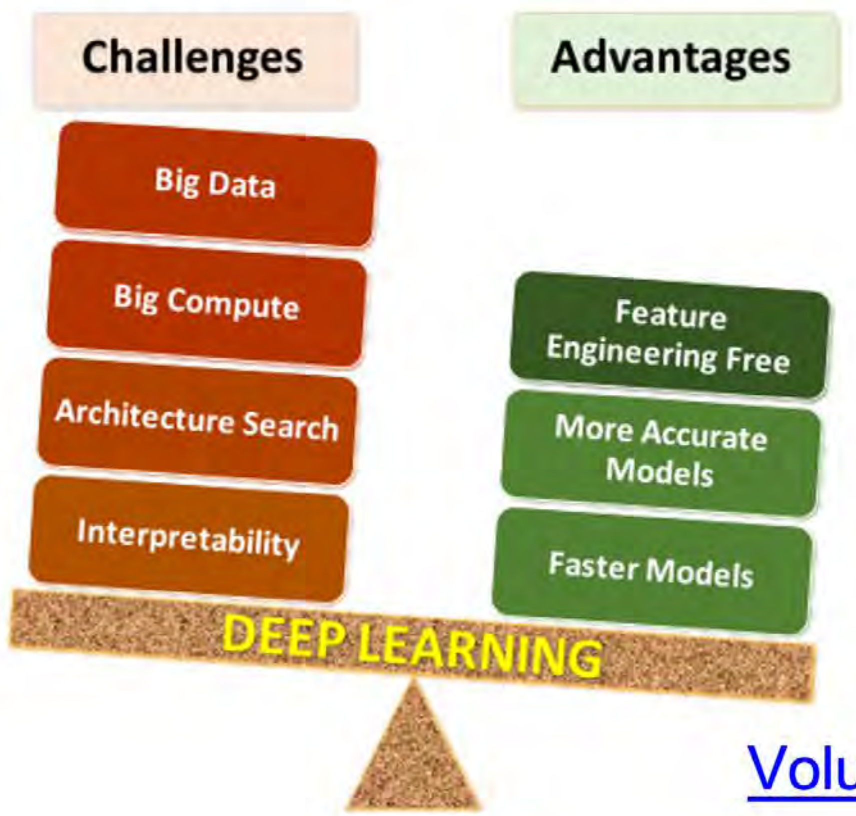
- Material Science and Design
- Climate Understanding
- Cosmology

MRS Communications Artificial Intelligence Prospective

Deep materials informatics: Applications of deep learning in materials science

Ankit Agrawal and **Alok Choudhary,** Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60201, USA
Address all correspondence to Ankit Agrawal at **ankitag@eecs.northwestern.edu**

Abstract

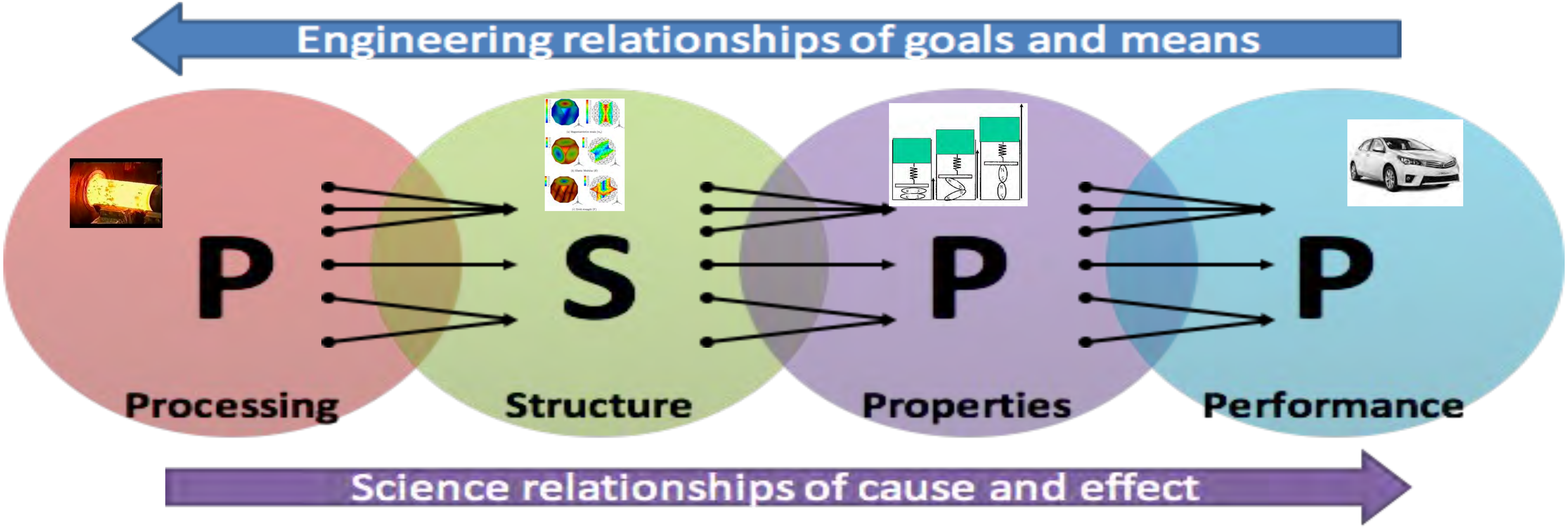


The growing application of data-driven analytics in materials science has led to the rise of materials informatics. Within the arena of data analytics, deep learning has emerged as a game-changing technique in the last few years, enabling numerous real-world applications, such as self-driving cars. In this paper, the authors present an overview of deep learning, its advantages, challenges, and recent applications on different types of materials data. The increasingly availability of materials databases and big data in general, along with groundbreaking advances in deep learning offers a lot of promise to accelerate the discovery, design, and deployment of next-generation materials.

Volume 9, Issue 3, September 2019, pp. 779-792

PSPP Relationships in Materials

Inverse Problem



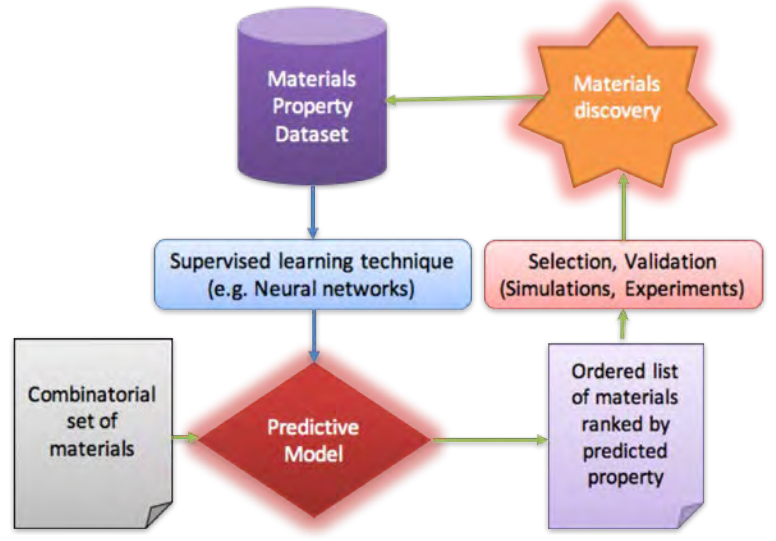
Single AI/ML step applicable to multiple design problems

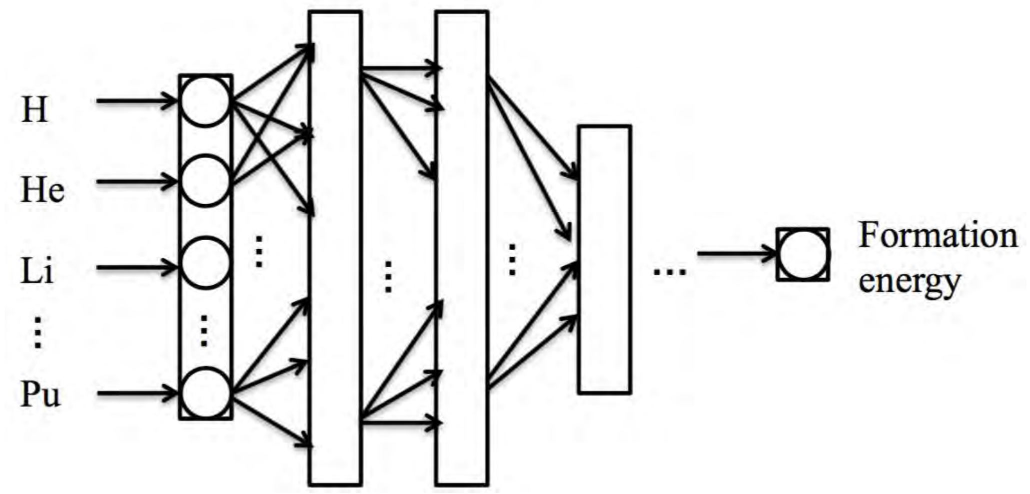
- <u>Data</u>
  - Hundreds of thousands of DFT calculations (e.g., OQMD)
  - <u>Composition-based models</u>
  - 145 attributes (stoichiometric/ elemental/electronic/ionic)
    - Mean Electronegativity
    - Bond Ionic Character… (148+ properties)
- <u>Structure-aware models</u>
  - Voronoi tessellations to capture local environment of atoms
- <u>Inverse models</u>
  - Stable compounds, metallic glasses, semiconductors?

Online Tool: http://info.eecs.northwestern.edu/FEpredictor

COMPOSITION-DERIVED ATTRIBUTES

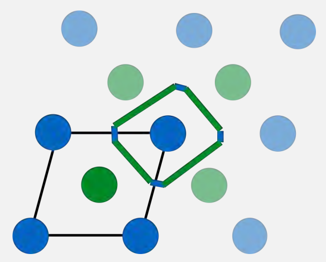
Attribute	*Value for Fe_2O_3*
Average atomic mass	0.4x55.845 + 0.6x15.999 = 31.94
Average column on periodic table	0.4x8 + 0.6x16 = 12.8
Average row on the periodic table	0.4x4 + 0.6x2 = 2.8
Maximum difference in atomic number	26 - 8 = 18
Average atomic number	0.4x26 + 0.6x8 = 15.2
Maximum difference in atomic radii (pm)	140 - 60 = 80
Average atomic radius	0.4x140 + 0.6x60 = 92.0
Maximum difference in electronegativity	3.44 - 1.83 = 1.61
Average electronegativity	0.4x3.44 + 0.6x1.83 = 2.474
Average number of s valence electrons	0.4x4 + 0.6x2 = 2.8
Average number of p valence electrons	0.4x0 + 0.6x4 = 2.4
Average number of d valence electrons	0.4x6 + 0.6x0 = 2.4
Average number of f valence electrons	0.4x0 + 0.6x0 = 0.0
s fraction of valence electrons	2.8 / (2.8+2.4+2.4+0.0) = 0.368
p fraction of valence electrons	2.4 / (2.8+2.4+2.4+0.0) = 0.316
d fraction of valence electrons	2.4 / (2.8+2.4+2.4+0.0) = 0.316
f fraction of valence electrons	0.0 / (2.8+2.4+2.4+0.0) = 0.0





AI/ML for Science – Materials Property Prediction

Density Functional Theory (Structure-aware models)



Data from Thousands of DFT Simulations (Unexploited knowledge base)

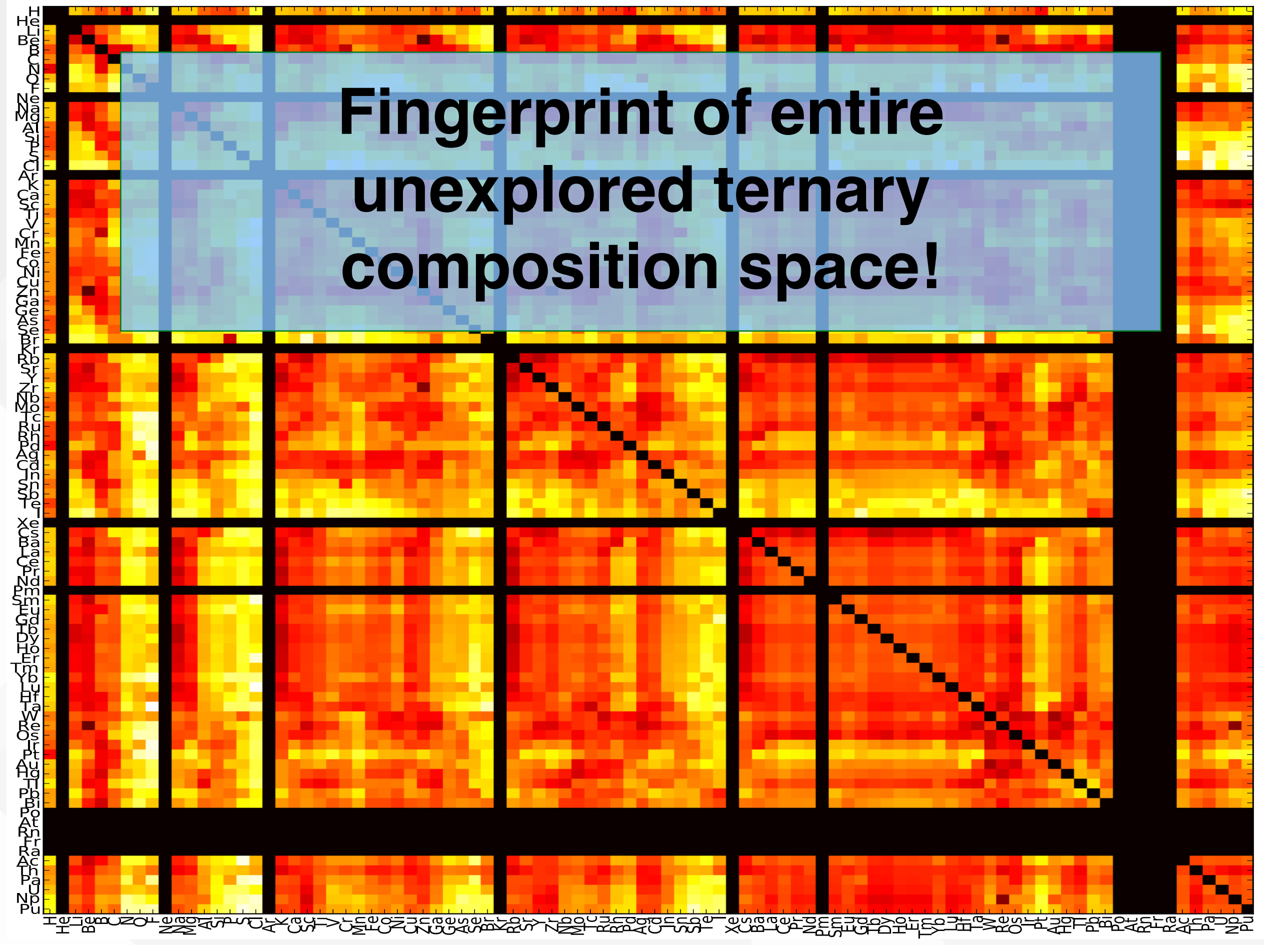
AI/ML - Predictive Modeling (Learn model for property of interest)

Insights (Quick, Prioritize)

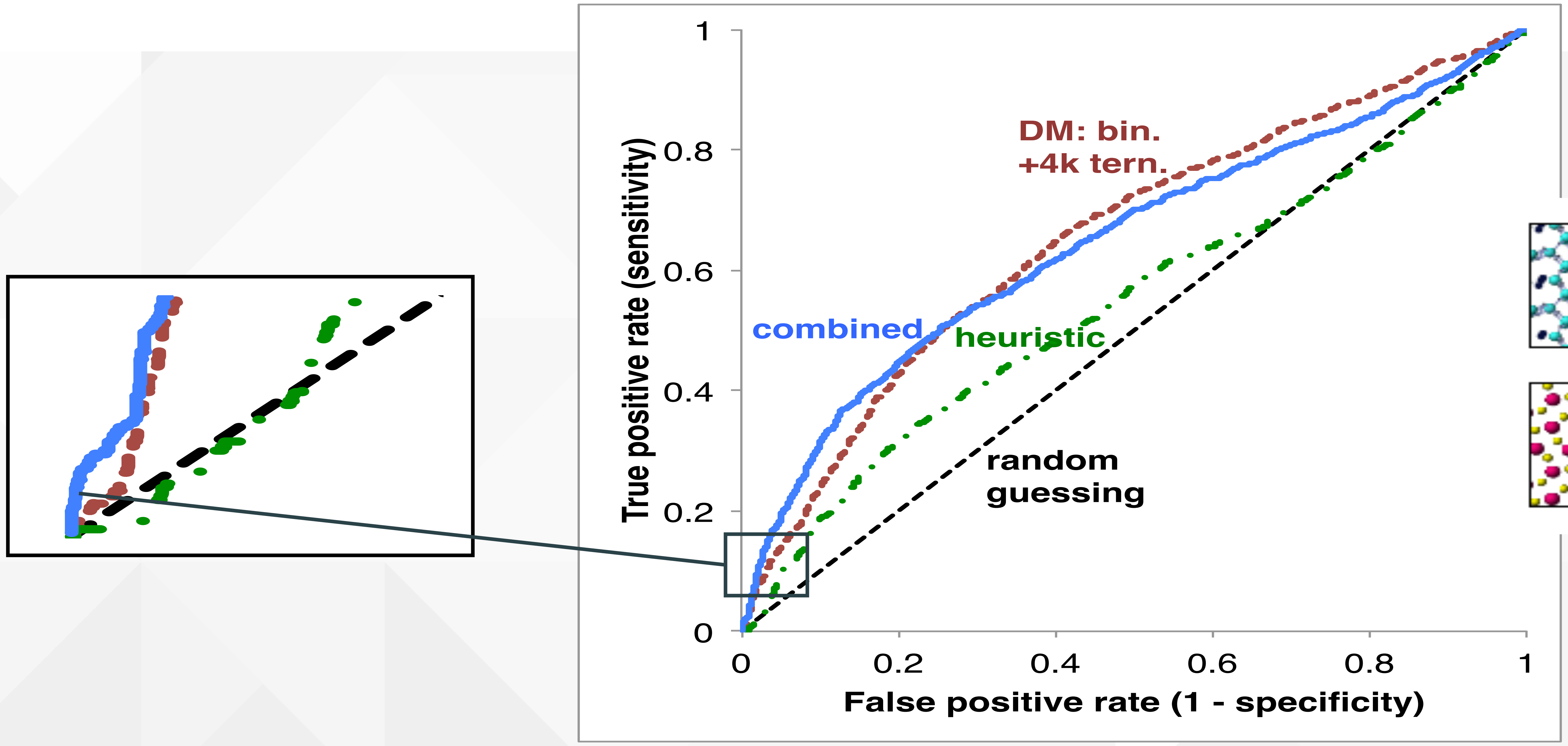
$

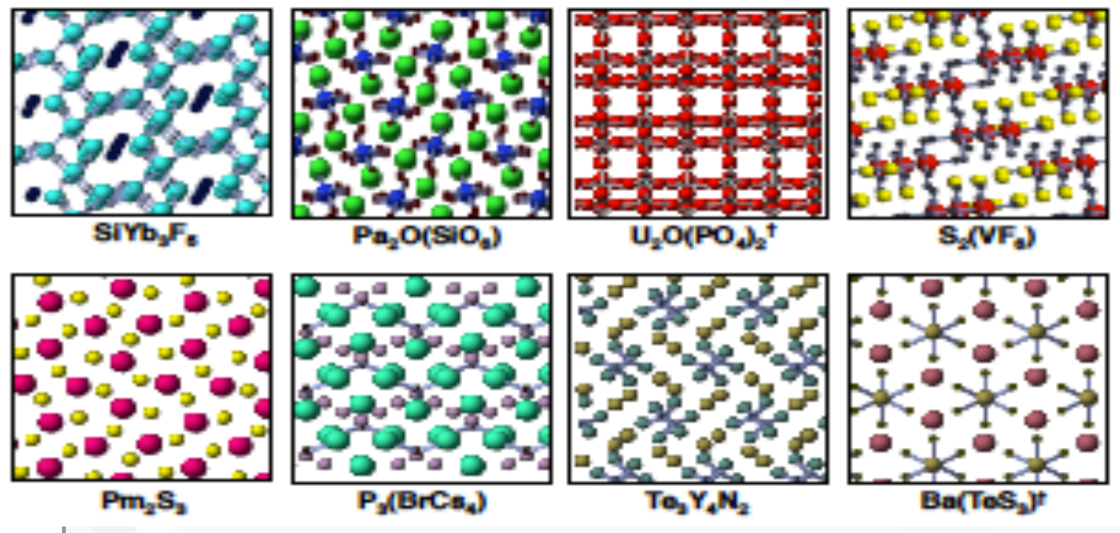
Experiments or Simulations

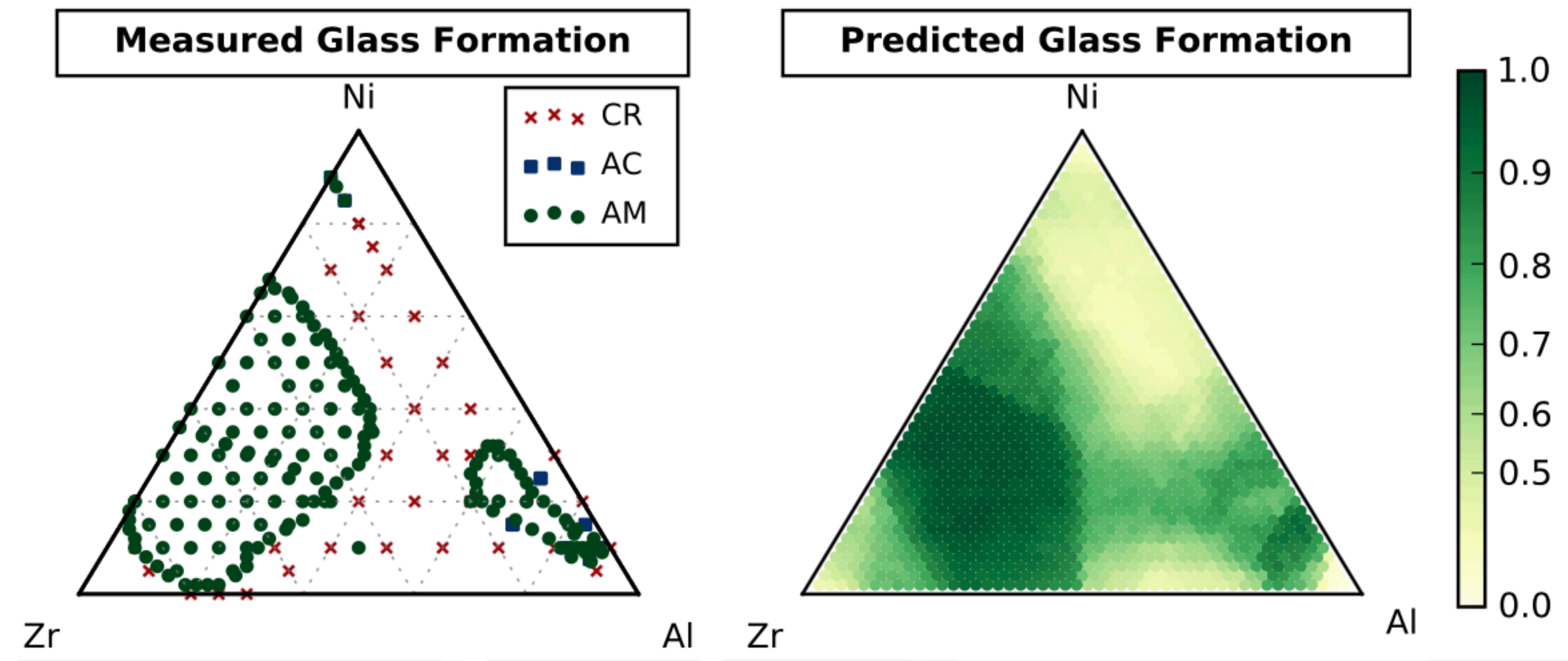
Semiconductors – E.g., Formation Energy



Inference Engine







New Semiconductors

Composition	E_g *(eV)*
$ScHg_4Cl_7$	1.26
$V_2Hg_3Cl_7$	1.16
Mn_6CCl_8	1.28
$Hf_4S_{11}Cl_2$	1.11
VCu_5Cl_9	1.19

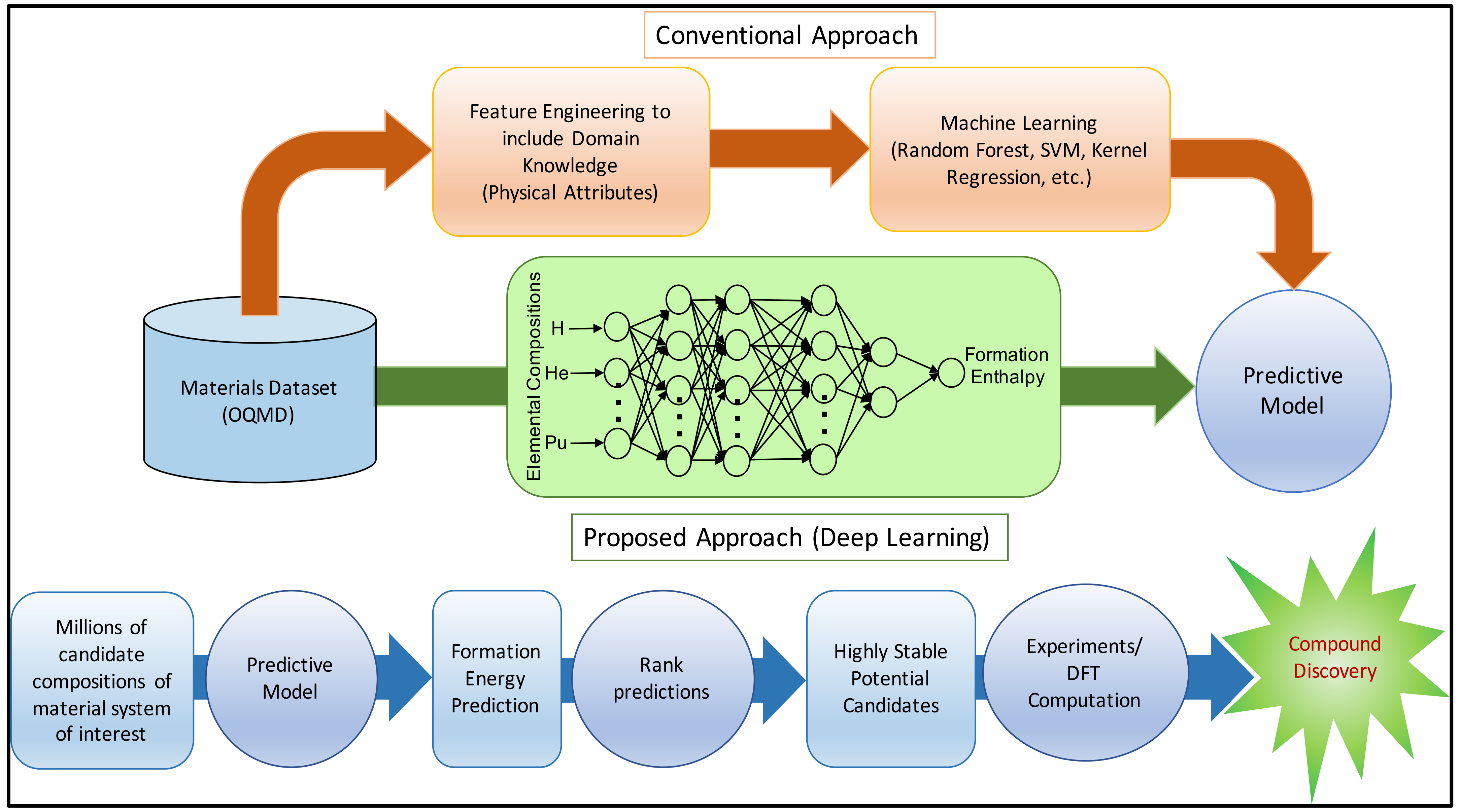
New metallic glasses

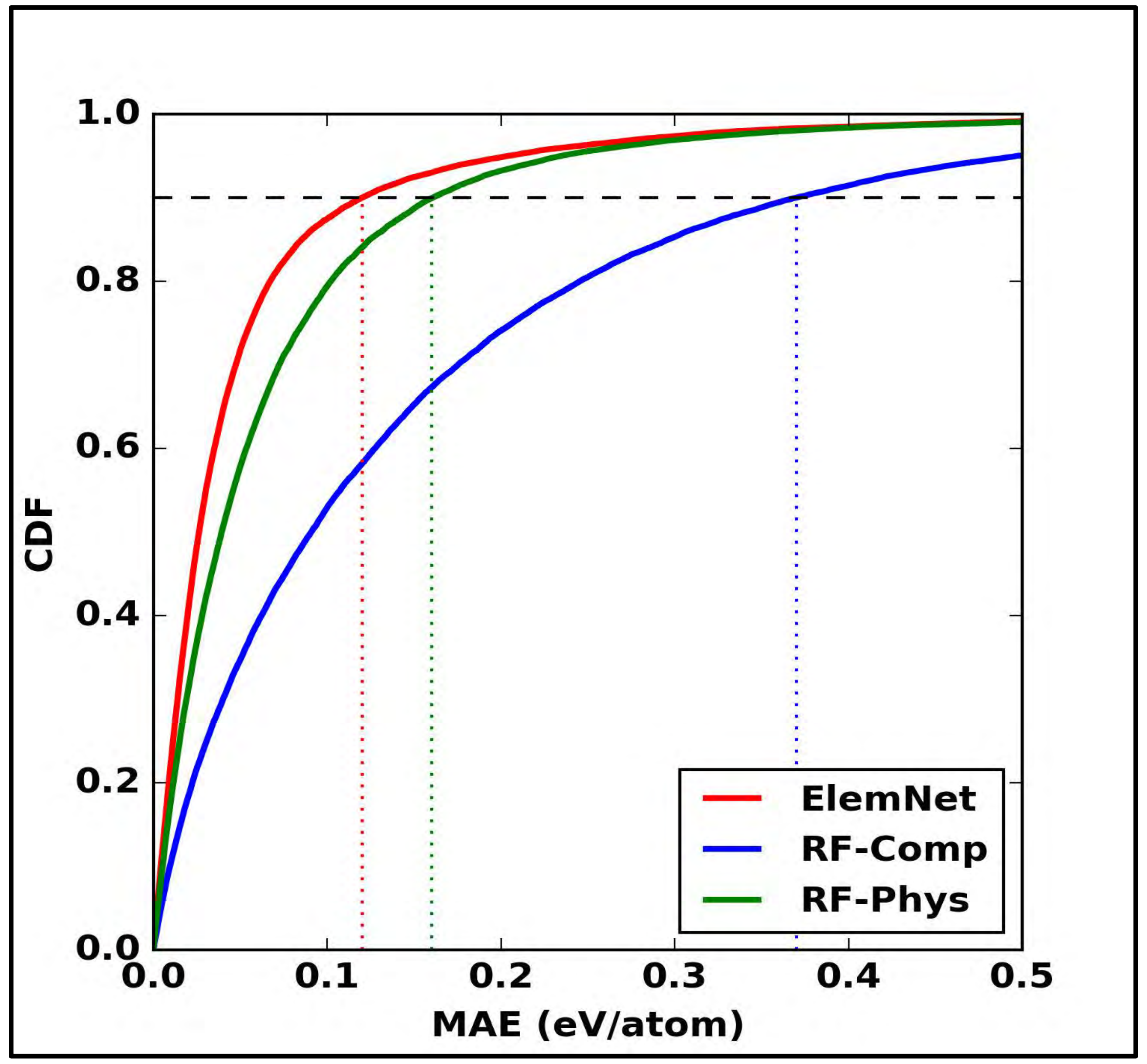
Alloy composition
$Zr_{0.38}Co_{0.24}Cu_{0.38}$
$V_{0.16}Ni_{0.64}B_{0.2}$
$Zr_{0.46}Cr_{0.36}Ni_{0.18}$
$Zr_{0.5}Fe_{0.38}W_{0.12}$

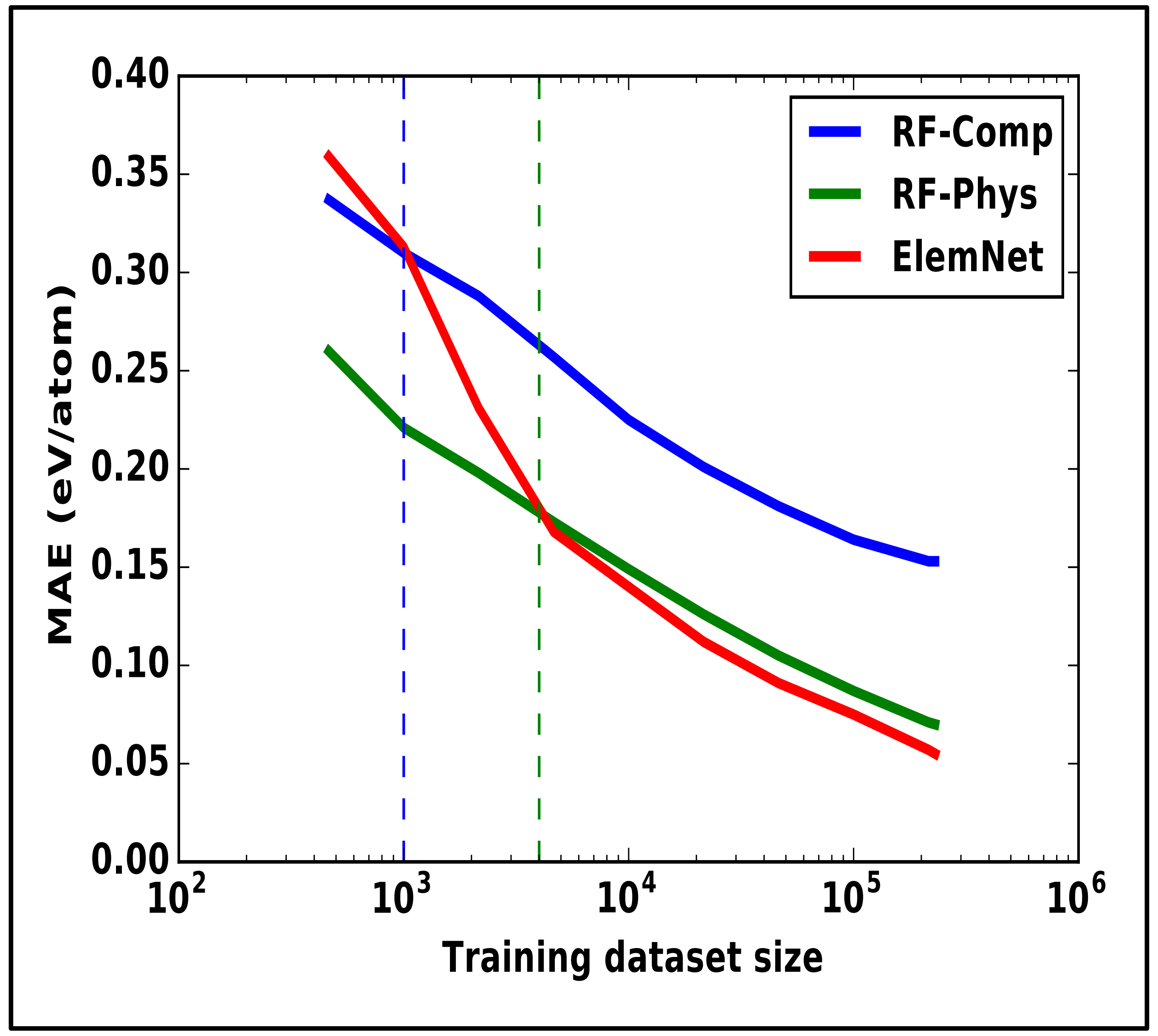
ElemNet: Learning Chemistry

From Only Element Composition

- ML models need domain knowledge
- **ElemNet** learns from elemental compositions only
- Captures the similarity and chemical interactions between different elements.
- Better accuracy at two order of magnitude faster rate than traditional ML model
- Fast and robust combinatorial screening in huge composition space of billions of compounds.







D. Jha, L. Ward, A. Paul, W.-keng Liao, A. Choudhary, C. Wolverton, and A. Agrawal, "ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition," Nature *Scientific Reports*, vol. 8, no. 17593, 2018.

Deep Transfer Learning for (Small) Experimental Datasets

Challenge

- Most materials datasets are *small*
- DFT vs experiment: Formation enthalpy MAE = ~0.08 eV/atom

Methodology

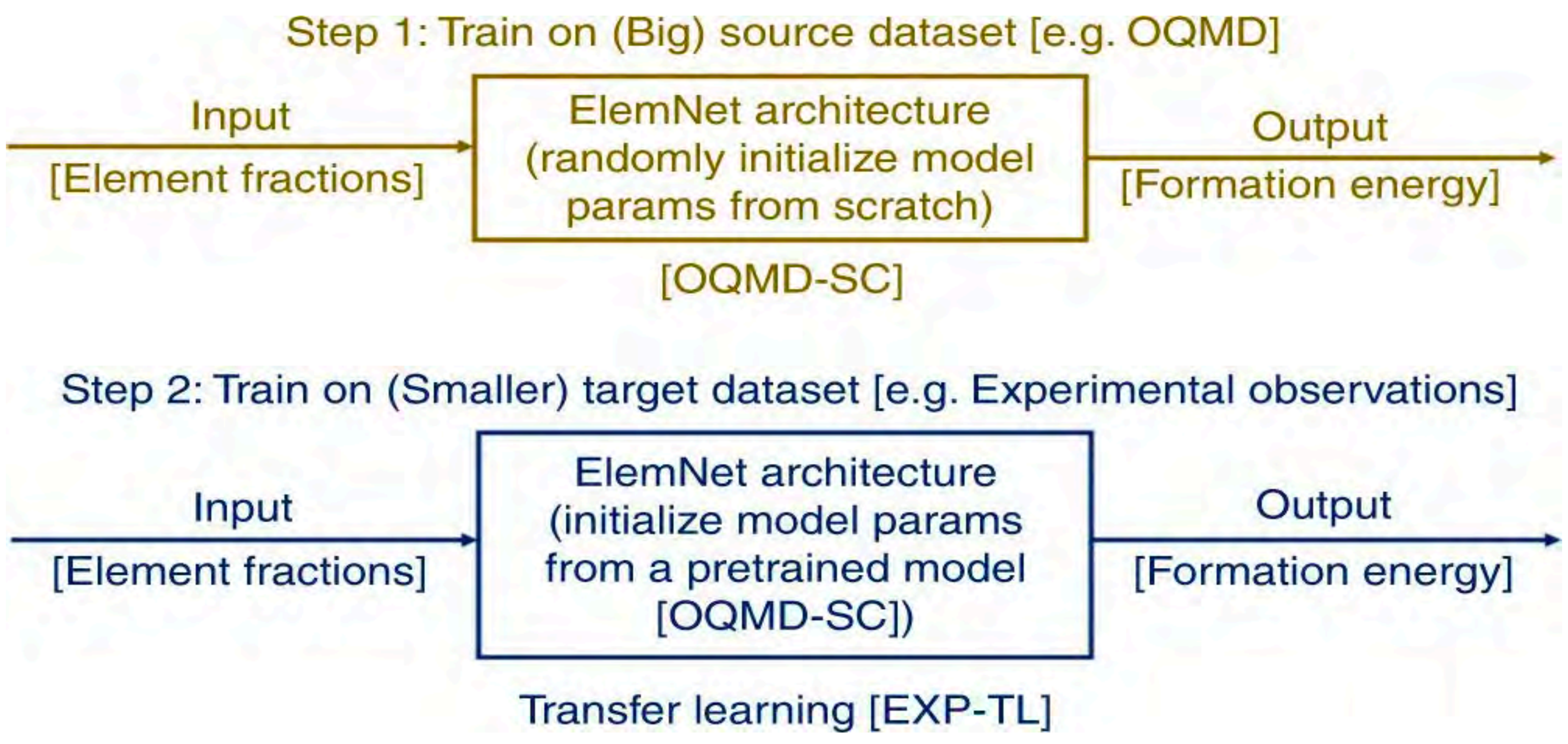
- *Deep transfer learning*
- Refine weights of a model pretrained on Simulation data

Datasets

- Source: OQMD
- Target: JARVIS, MP, Experimental

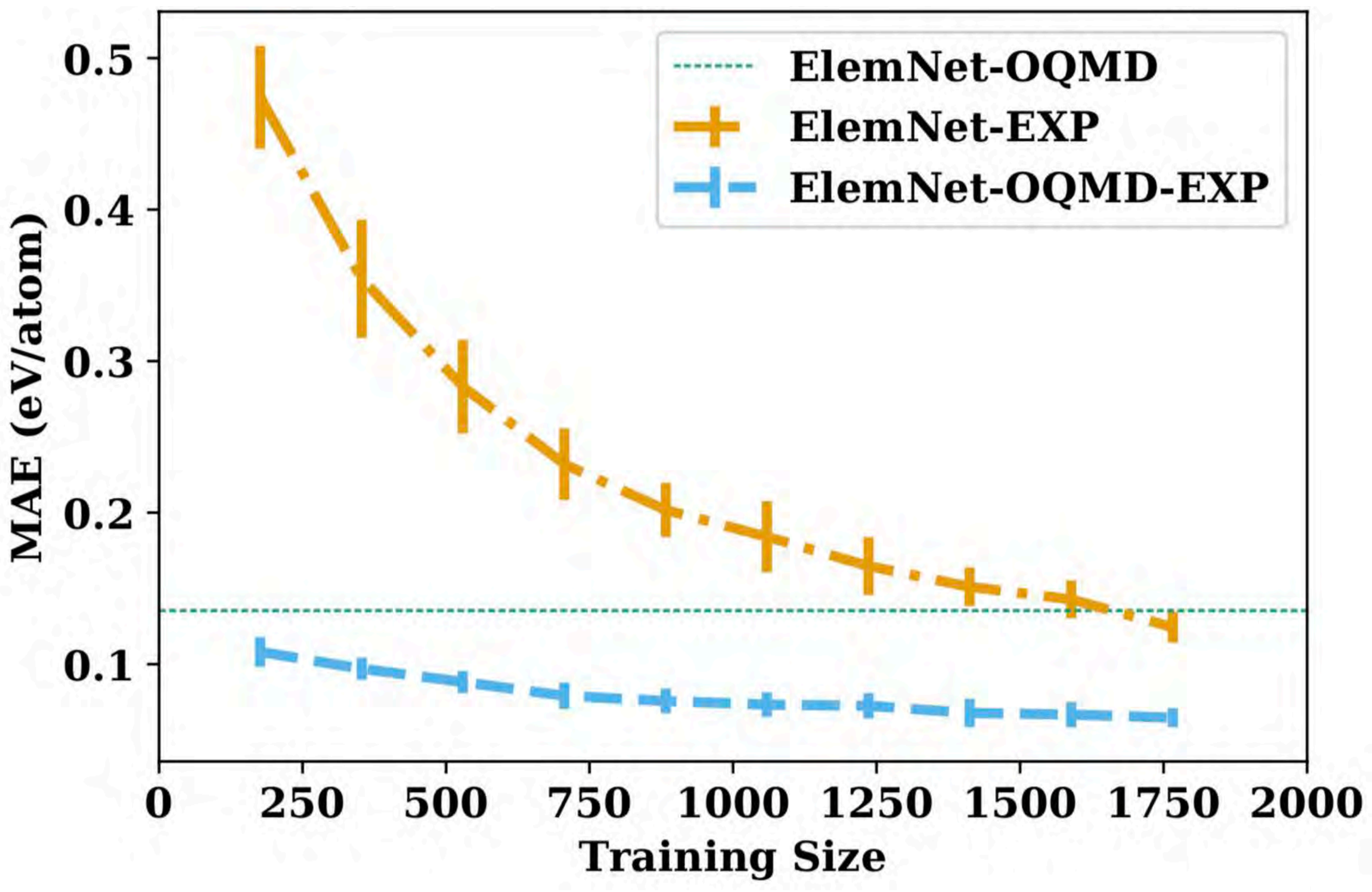
Results

- TL model > Training from scratch
- Up to 58% reduction in MAE on small (<2K) experimental data
- MAE_{FE} = ~0.06 eV/atom



Dataset	Size	Training from Scratch	Transfer Learning
		MAE (eV/atom)	MAE (eV/atom)
OQMD	341,000	0.0437	-
JARVIS	11,050	0.0568	**0.0312**
Materials Project	23,641	0.0327	**0.0247**
Experimental	1,963	0.1460	**0.0608**

Jha et al., *Nature Communications,* 2019



Example: Industrial Materials Design

A complex and expensive work-flow

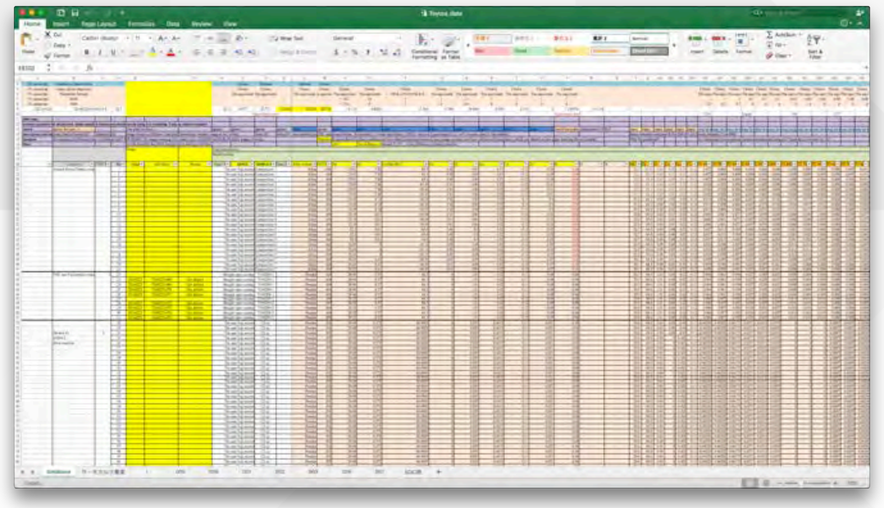
Samples → 1 → Rapid cooling → 2 → Molding → 3 → Hot deform → 4 → Heat treatment → 6 → Property2

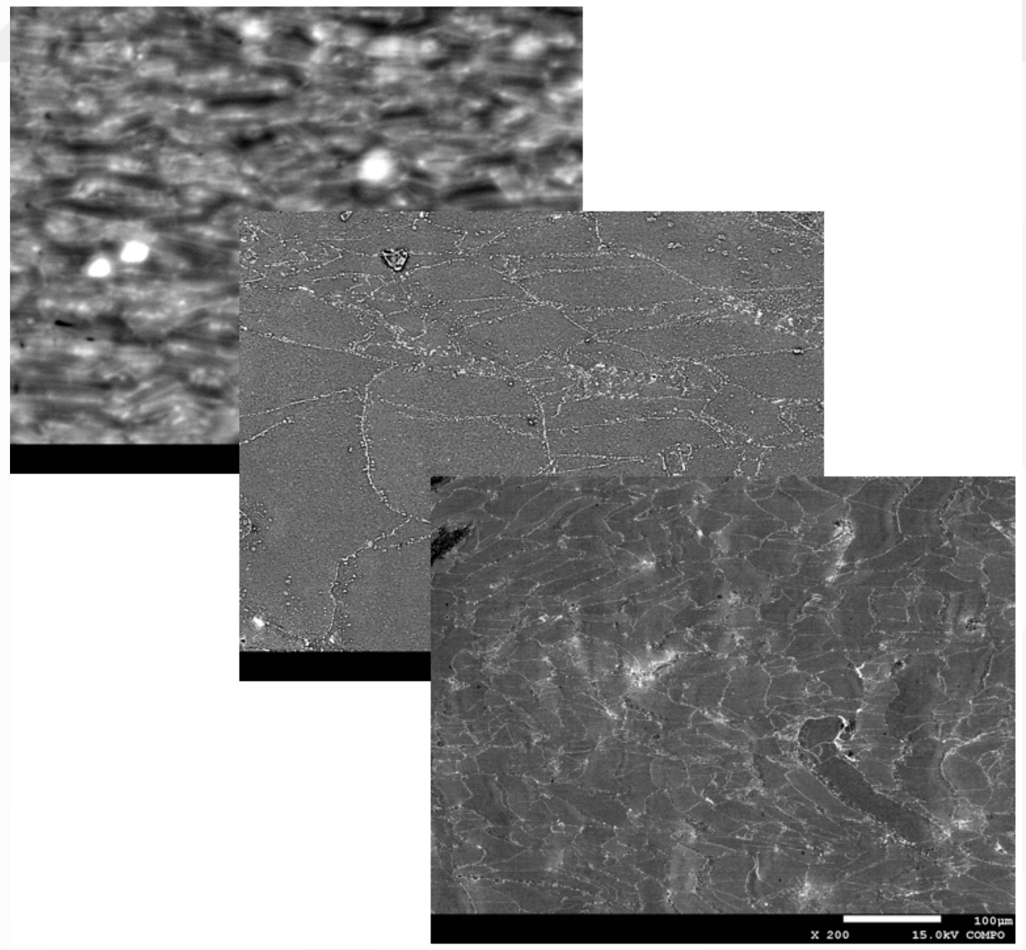
3 → Image1 (SEM)

5 → Property1

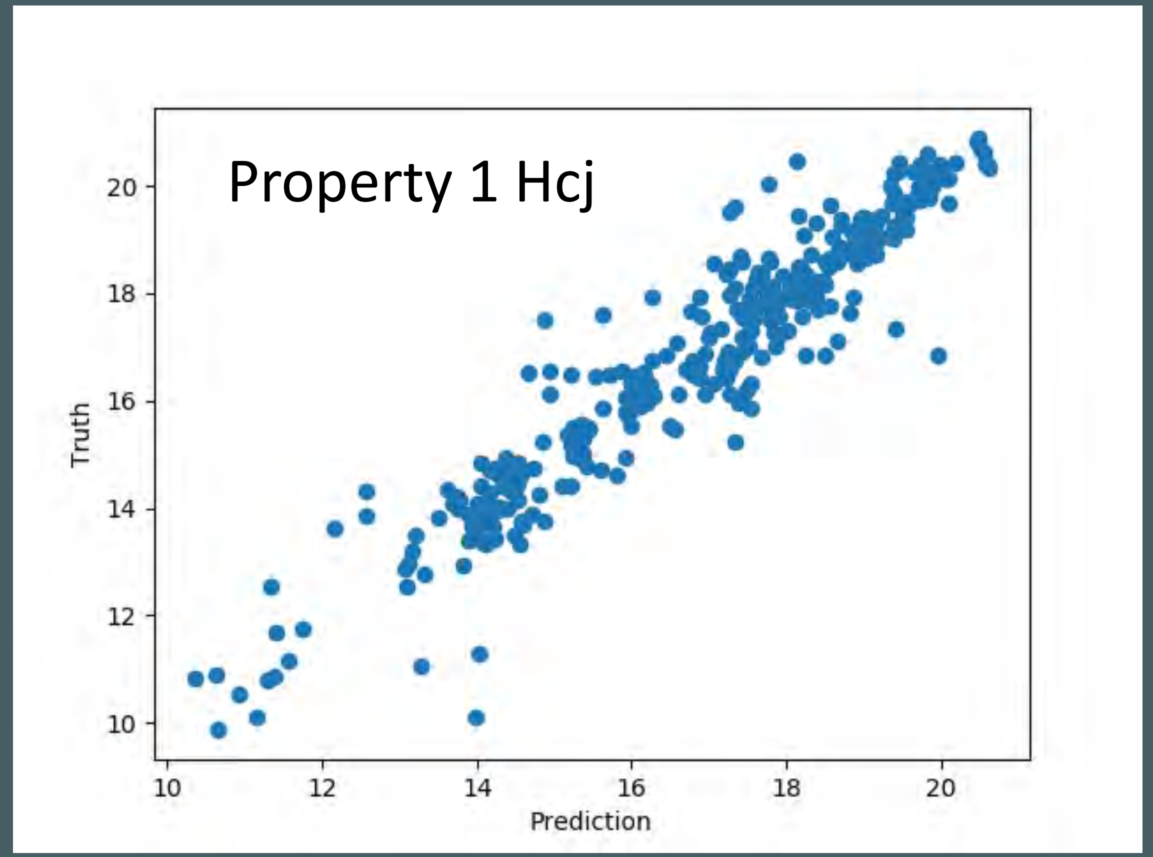
Image2 (SEM)

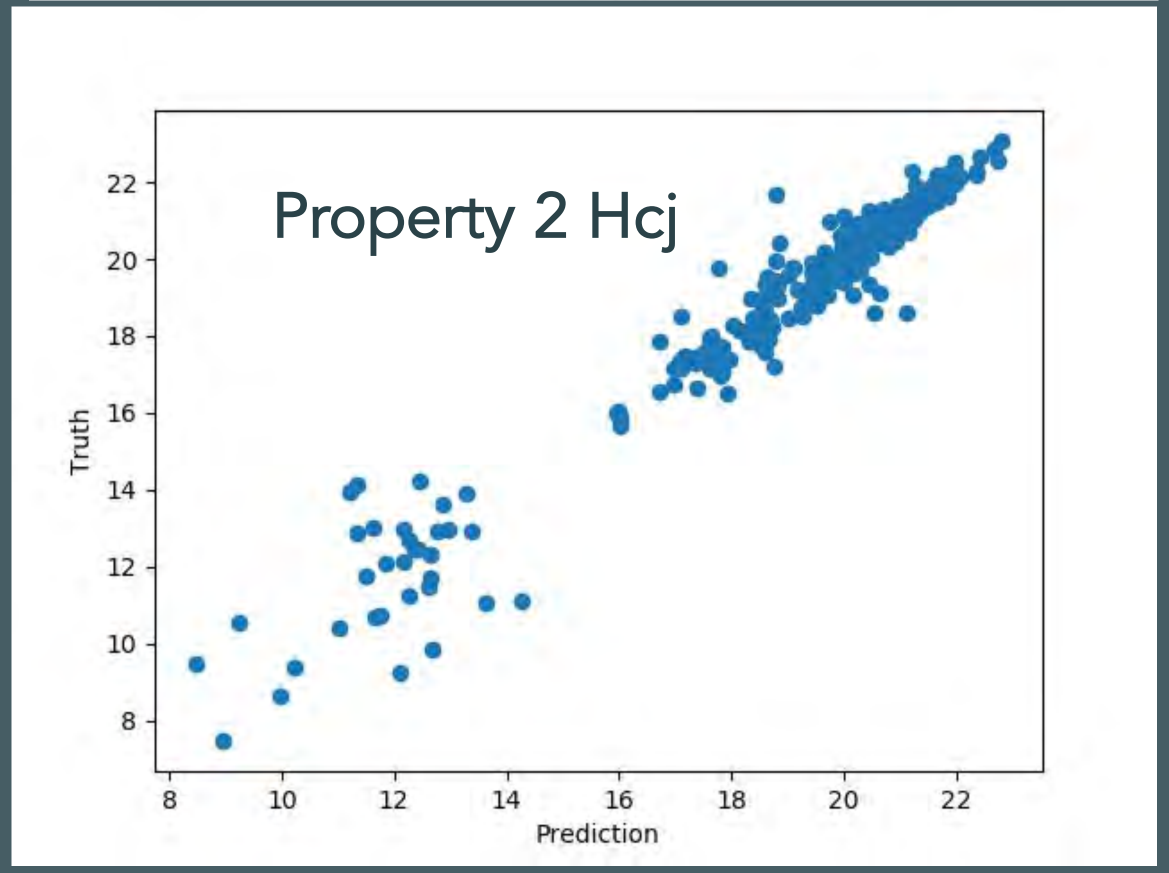
Models 1- 6

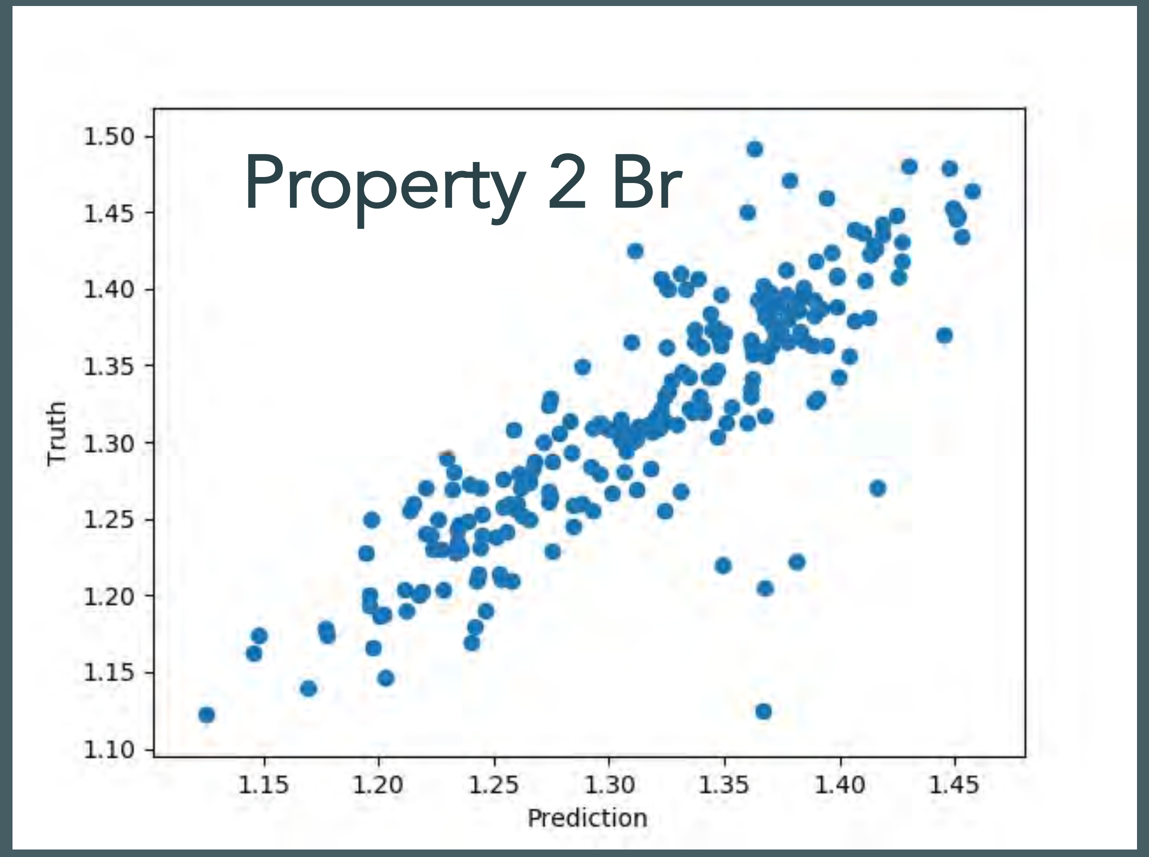




- Prediction of properties of samples under certain processing conditions



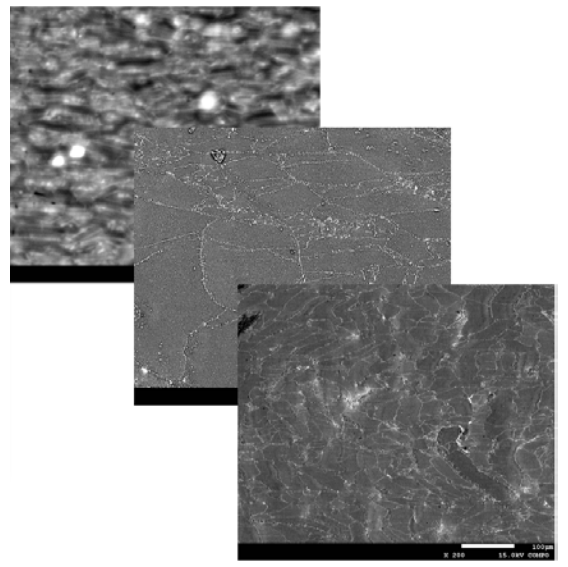




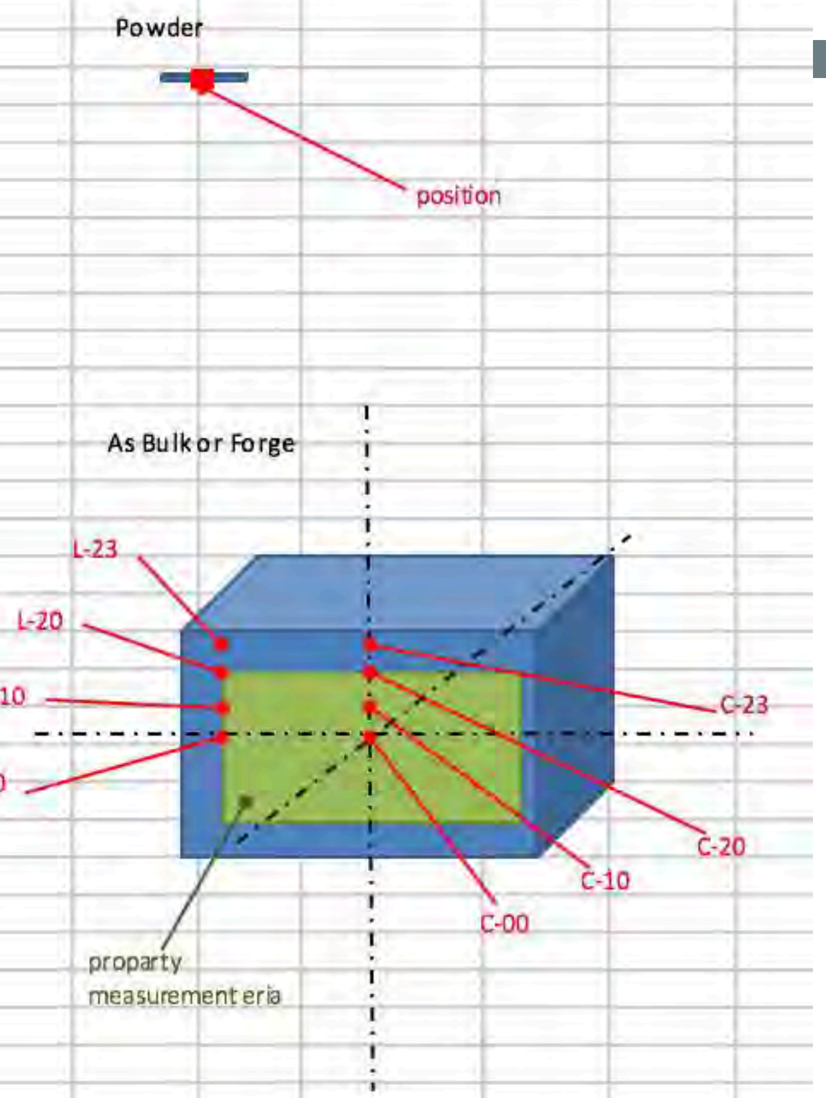
Property Prediction – numerical data

Image data

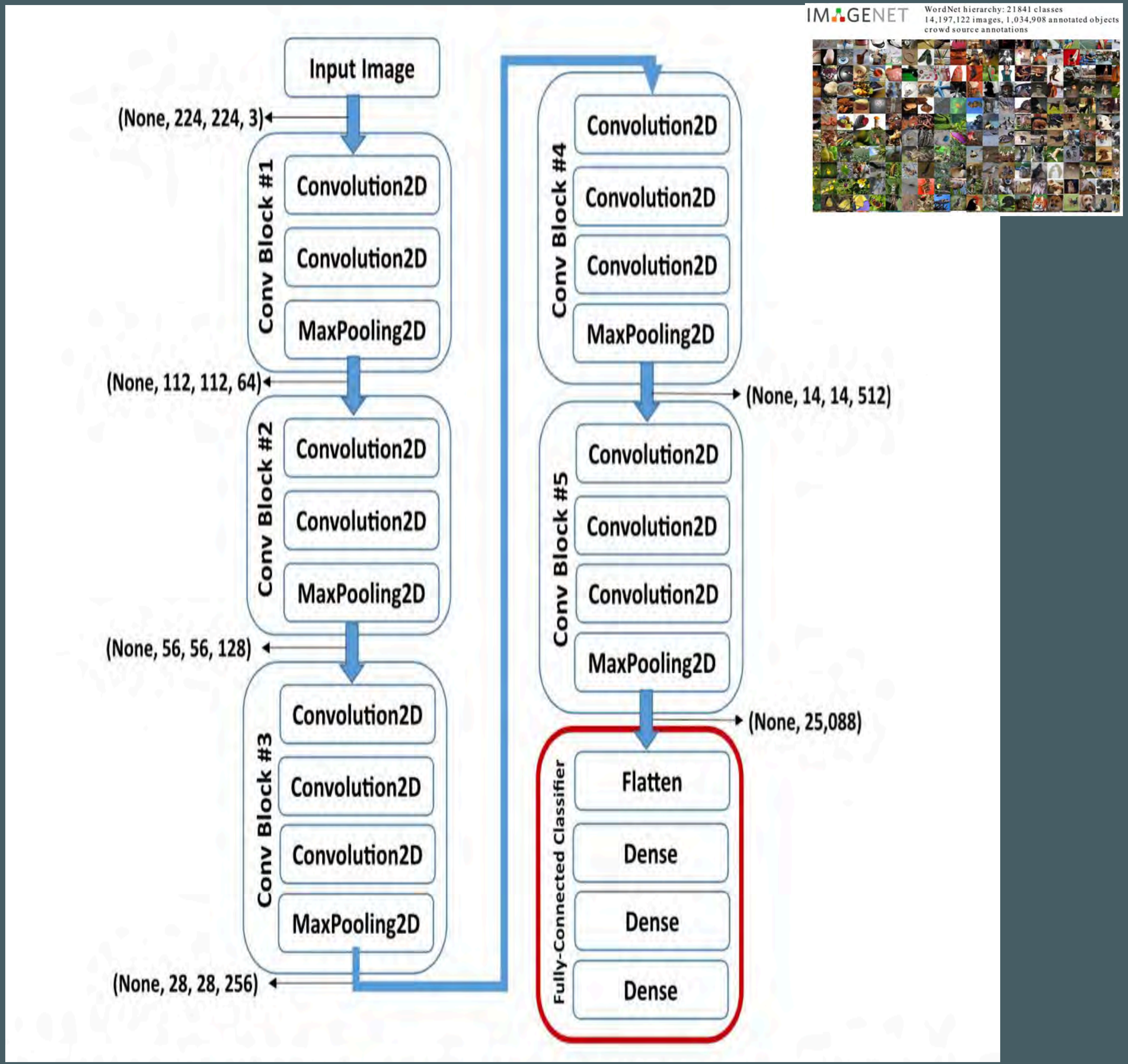
- Two image modes: COMPO, SEI
- Two targets: powder, as bulk or forge
- eight positions: C00, C10, C20, C23, L00, L10, L20, L23
- magnifications: x200, x1000 and x30000



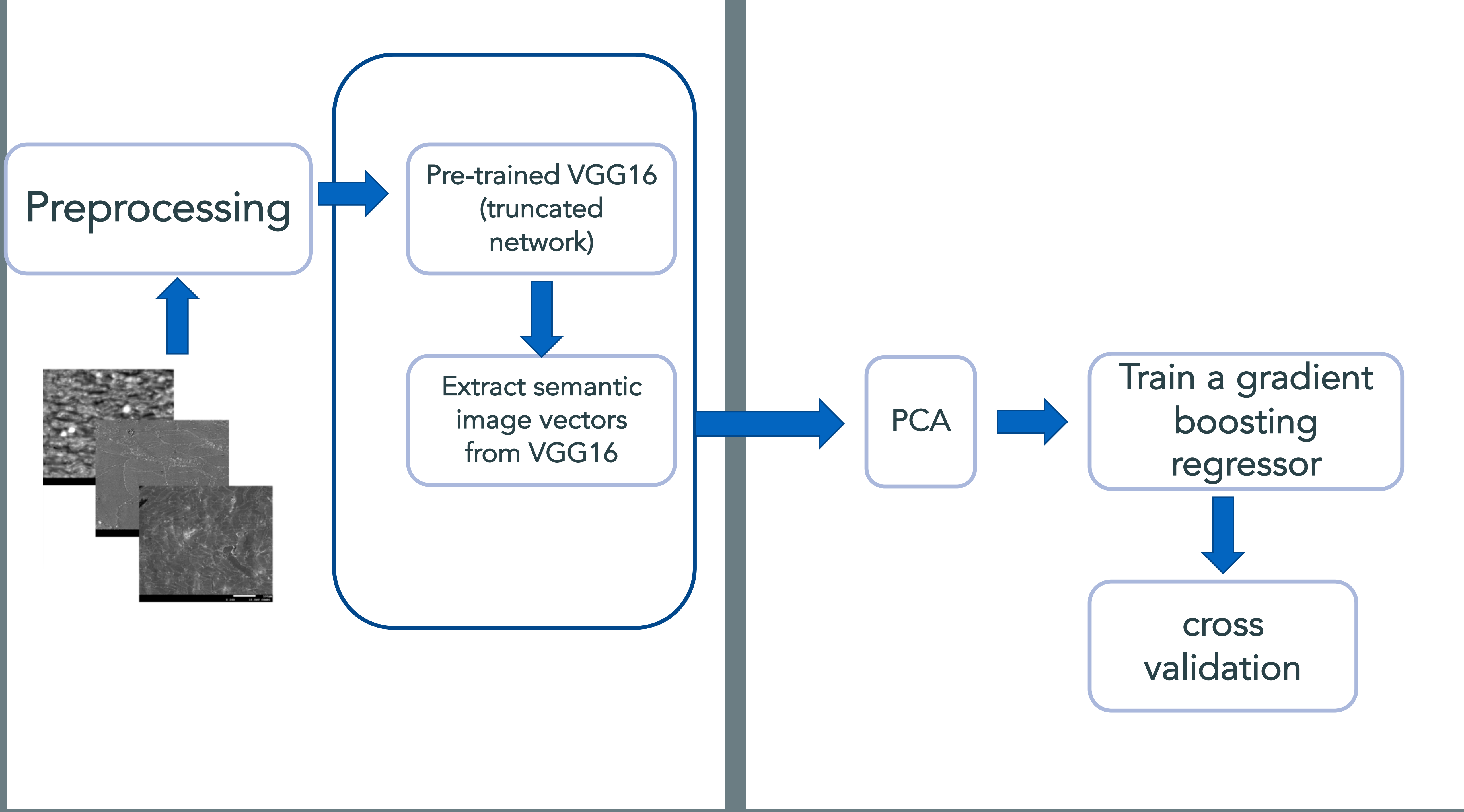
folder	file name	machine	mode	target	position (fig.)	magnification
T160223	T160223-068	SEM	SE	powder N1	cross section/ part of length	×2000
	T160223-069		COMPO		cross section/ part of length	×2000
	T160223-070	SEM	SE	powder N2	cross section/ part of length	×2000
	T160223-071		COMPO		cross section/ part of length	×2000
	T160223-072	SEM	SE	powder N3	cross section/ part of length	×2000
	T160223-073		COMPO		cross section/ part of length	×2000
	T160223-074	SEM	SE	powder N4	cross section/ part of length	×2000
	T160223-075		COMPO		cross section/ part of length	×2000
B150831	B150831-08-C-00-x30k-01	SEM	COMPO	As forge		
	B150831-08-C-00-x30k-02		SE		cross section/ C-00	×30000
	B150831-08-C-00-x30k-03	SEM	COMPO	As forge		
	B150831-08-C-00-x30k-04		SE		cross section/ C-00	×30000
	B150831-08-C-00-x200-01	SEM	COMPO	As forge		
	B150831-08-C-00-x200-02		SE		cross section/ C-00	×200
	B150831-08-C-00-x1000-01	SEM	COMPO	As forge		
	B150831-08-C-00-x1000-02		SE		cross section/ C-00	×1000
	B150831-08-C-10-x30k-01	SEM	COMPO	As forge		
	B150831-08-C-10-x30k-02		SE		cross section/ C-10	×30000
	B150831-08-C-10-x30k-03	SEM	COMPO	As forge		
	B150831-08-C-10-x30k-04		SE		cross section/ C-10	×30000
	B150831-08-C-10-x200-01	SEM	COMPO	As forge		
	B150831-08-C-10-x200-02		SE		cross section/ C-10	×200
	B150831-08-C-10-x1000-01	SEM	COMPO	As forge		
	B150831-08-C-10-x1000-02		SE		cross section/ C-10	×1000
	B150831-08-C-20-x30k-01	SEM	COMPO	As forge		
	B150831-08-C-20-x30k-02		SE		cross section/ C-20	×30000
	B150831-08-C-20-x30k-03	SEM	COMPO	As forge		
	B150831-08-C-20-x30k-04		SE		cross section/ C-20	×30000
	B150831-08-C-20-x200-01	SEM	COMPO	As forge		
	B150831-08-C-20-x200-02		SE		cross section/ C-20	×200
	B150831-08-C-20-x1000-01	SEM	COMPO	As forge		
	B150831-08-C-20-x1000-02		SE		cross section/ C-20	×1000
	B150831-08-L-00-x200-01	SEM	COMPO	As forge		
	B150831-08-L-00-x200-02		SE		cross section/ L-00	×200
	B150831-08-L-10-x200-01	SEM	COMPO	As forge		
	B150831-08-L-10-x200-02		SE		cross section/ L-10	×200
	B150831-08-L-20-x200-01	SEM	COMPO	As forge		
	B150831-08-L-20-x200-02		SE		cross section/ L-20	×200
	B150831-08-L-23-x200-01	SEM	COMPO	As forge		
	B150831-08-L-23-x200-02		SE		cross section/ L-23	×200
	or later same with abave					

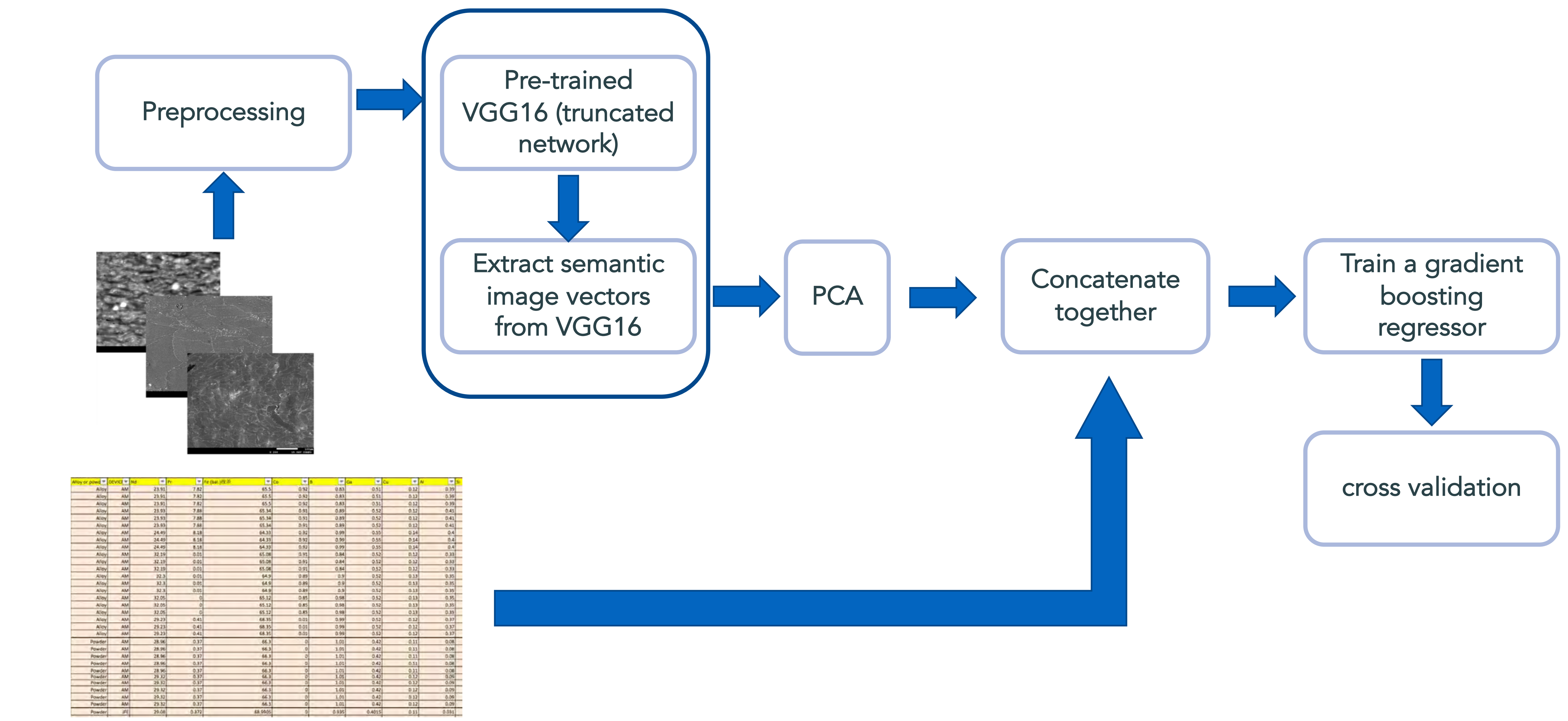


Deep (transfer) learning*



* What DL enables - Not enough Experimental data to learn





Combined Model

Impact!

- Tackle Complex Workflow
  - Many teams - Each needs expertise, resources and access
  - Involves Experiments, simulations, Instruments and ML
- Cost savings, Faster Exploration
  - E.g., 2 out of 8 image orientations have predictive value => significant reduction in (1) instruments (2) time, (3) sample materials
  - Fewer and relevant experiments - avoid back-end processing steps for not-so-promising candidates
  - Millions of $$$
- Discovery and Design acceleration
  - Explore and Discover most promising and high performing materials faster

Illustrative Publications - AI in Materials

- **Forward PSPP models (property prediction)**
  - Steels [IMMI 2014, CIKM 2016, IJF 2018, DSAA 2019]
  - Crystalline stability [PRB 2014, npjCM 2016, ICDM 2016, DL-KDD 2016, PRB 2017, SciRep 2018, KDD 2019, NatureComm 2019]
  - Band gap and glass forming ability prediction [npjCM 2016]
  - Bulk modulus prediction [RSC Adv 2016]
  - Seebeck coefficient prediction [JCompChem 2018]
  - Multi-scale localization/homogenization [IMMI 2015, IMMI 2017, CMS 2018, ActaMat 2019, IJCNN 2019]
  - Chemical properties prediction [NIPS MLMM 2018, IJCNN 2019, Molecular Informatics 2019]
- **Inverse PSPP models (optimization/discovery)**
  - Stable compounds [PRB 2014]
  - Magnetostrictive materials [Nature Scientific Reports 2015, AIAA 2018]
  - Semiconductors and metallic glasses [npjCM 2016]
  - Microstructure design (GAN) [JMD 2018]
  - Titanium aircraft panels [CMS 2019]
- **Structure characterization**
  - EBSD Indexing [BigData-ASH 2016, M&M 2018]
  - Crack detection in macroscale images [CBM 2017, IJTTE 2018]
  - XRD analysis for phase detection [IJCNN 2019]
  - Plastic deformation identification [IJCNN 2019]

MRS Communications

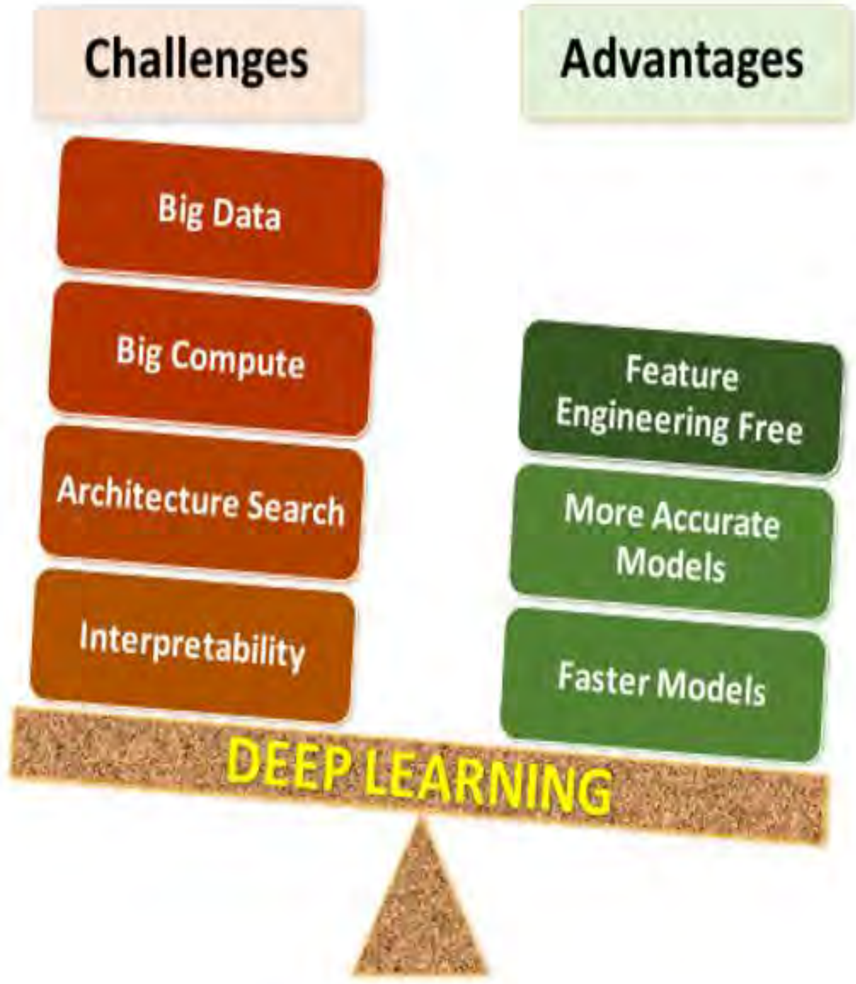
Artificial Intelligence Prospective

Deep materials informatics: Applications of deep learning in materials science

Ankit Agrawal and **Alok Choudhary,** Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60201, USA

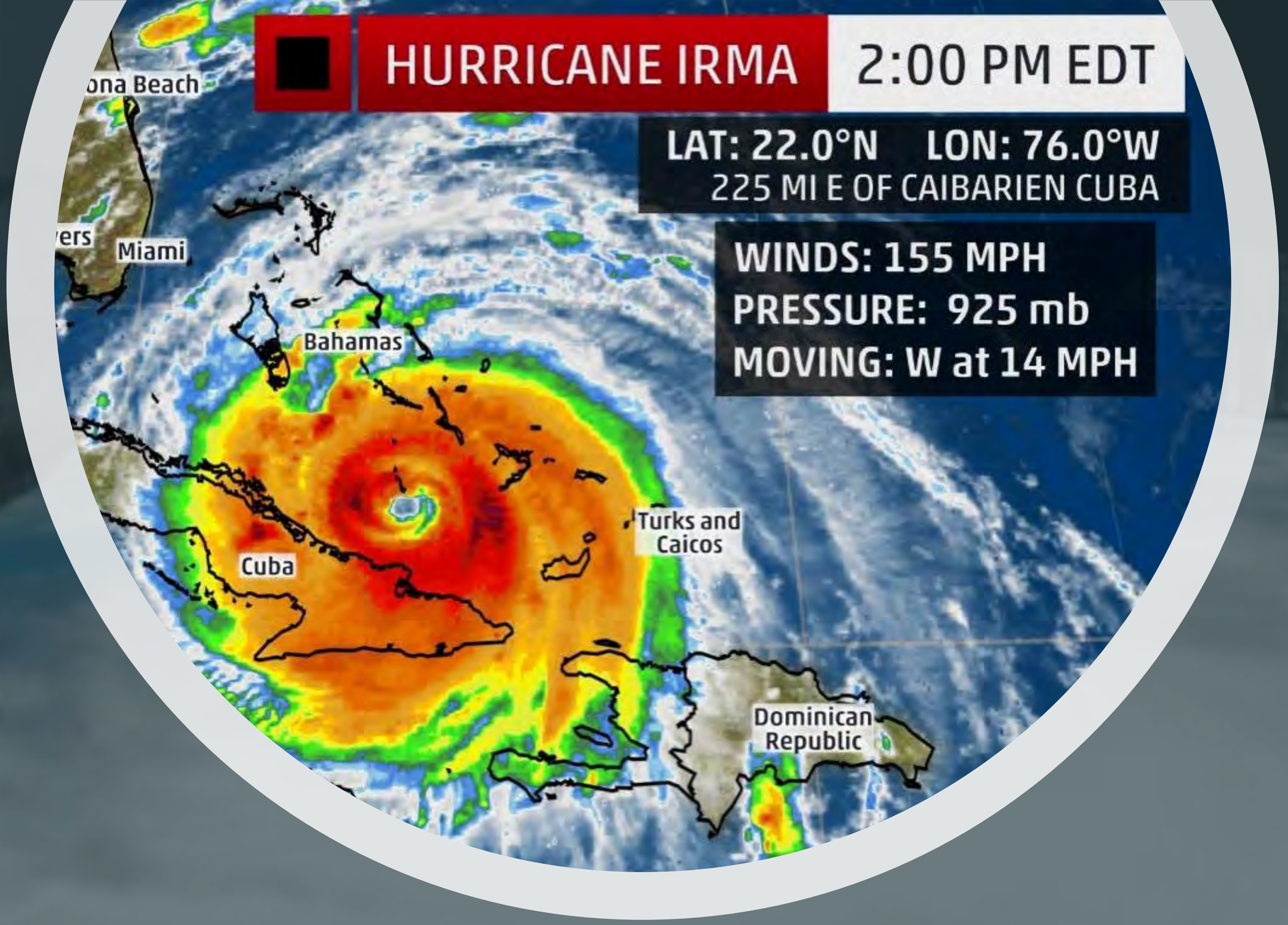
Address all correspondence to Ankit Agrawal at **ankitag@eecs.northwestern.edu**

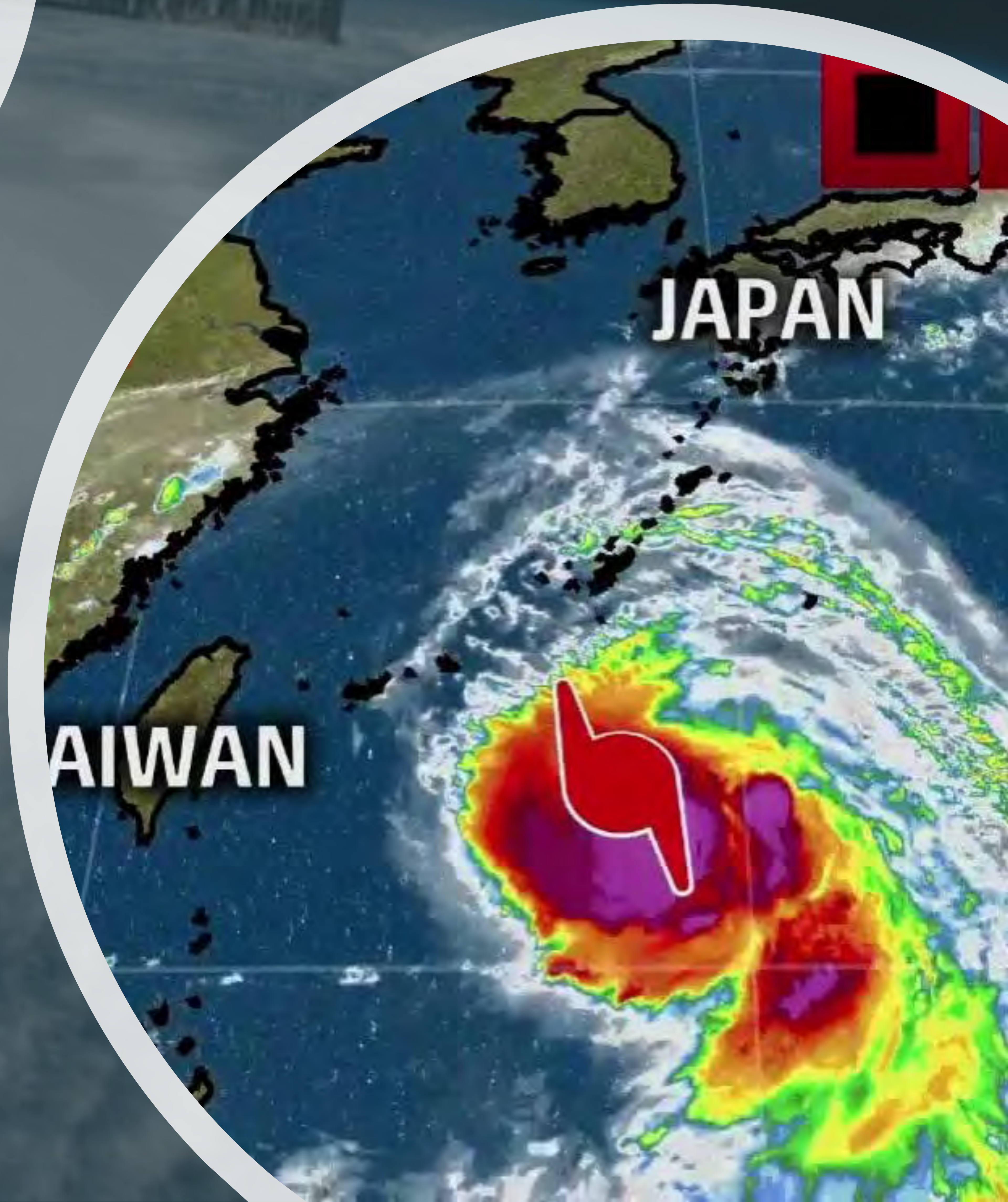
Abstract



The growing application of data-driven analytics in materials science has led to the rise of materials informatics. Within the arena of data analytics, deep learning has emerged as a game-changing technique in the last few years, enabling numerous real-world applications, such as self-driving cars. In this paper, the authors present an overview of deep learning, its advantages, challenges, and recent applications on different types of materials data. The increasingly availability of materials databases and big data in general, along with groundbreaking advances in deep learning offers a lot of promise to accelerate the discovery, design, and deployment of next-generation materials.

http://cucis.ece.northwestern.edu/publications/





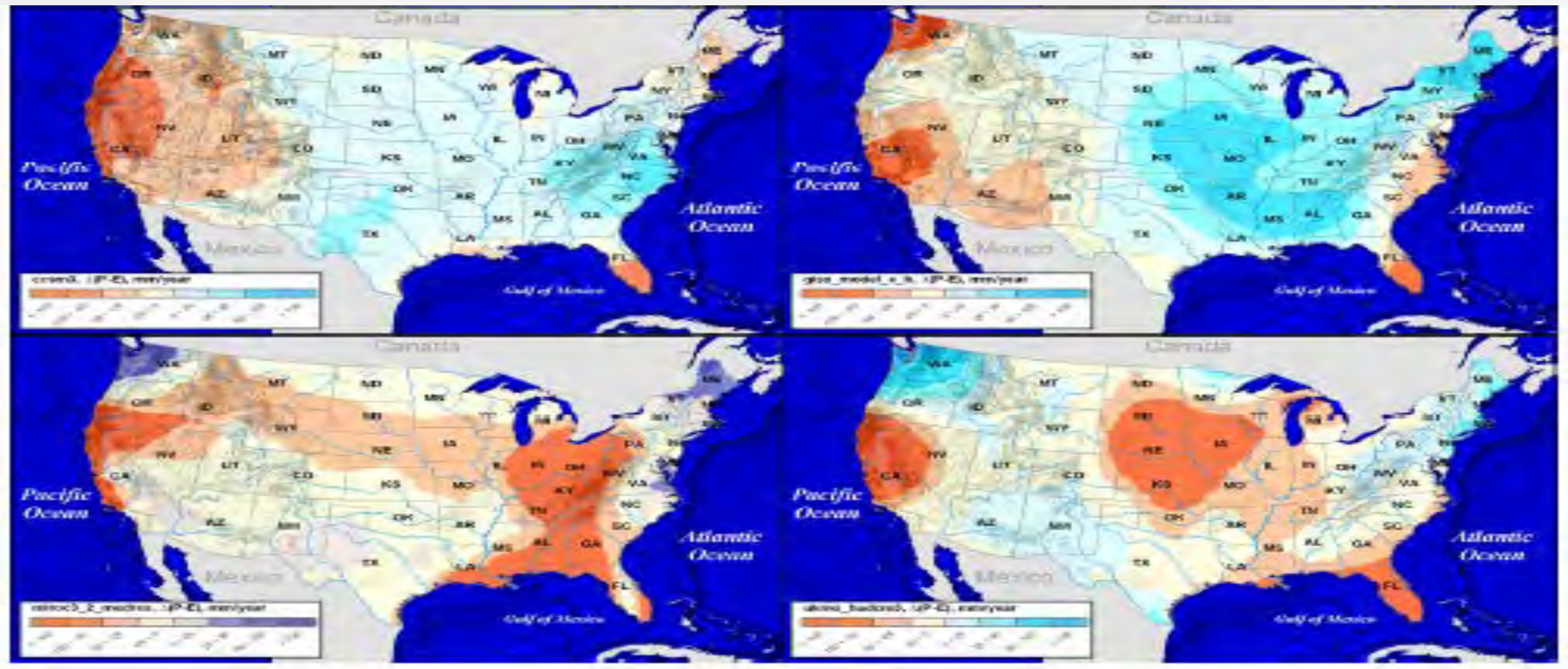
Understanding Climate Change

Limitation of Model Based Approaches

- Physics based models are essential but Limited
- Relatively reliable predictions at global scale for ancillary variables such as temperature
- Least reliable predictions for variables that are crucial for impact assessment such as regional precipitation

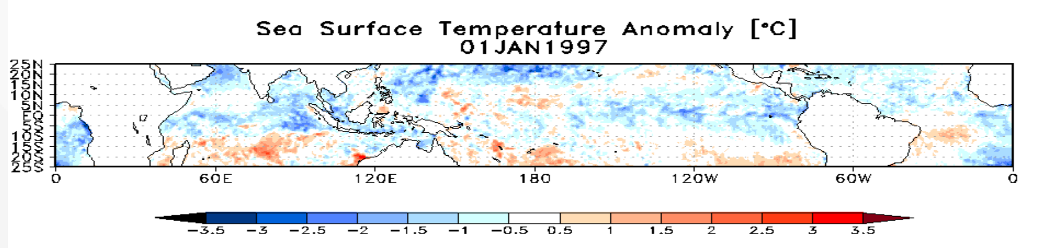
"The sad truth of climate science is that the most crucial information is the least reliable"
(Nature, 2010)

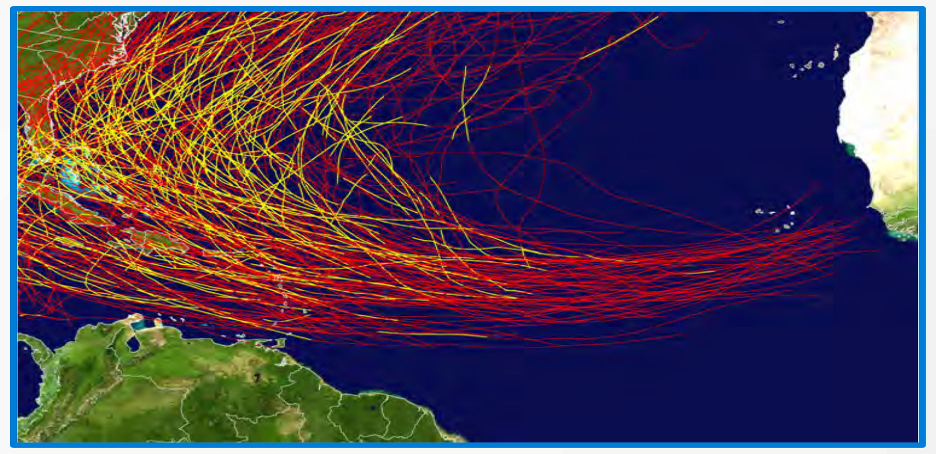
Disagreement between IPCC models

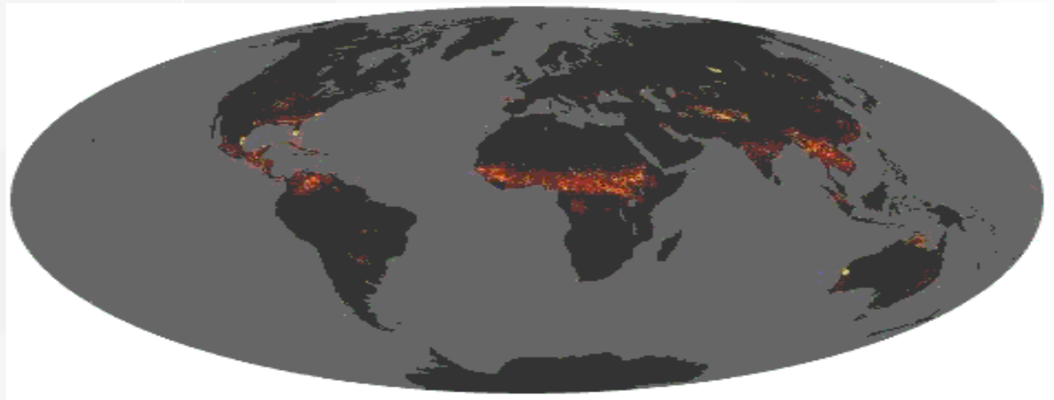


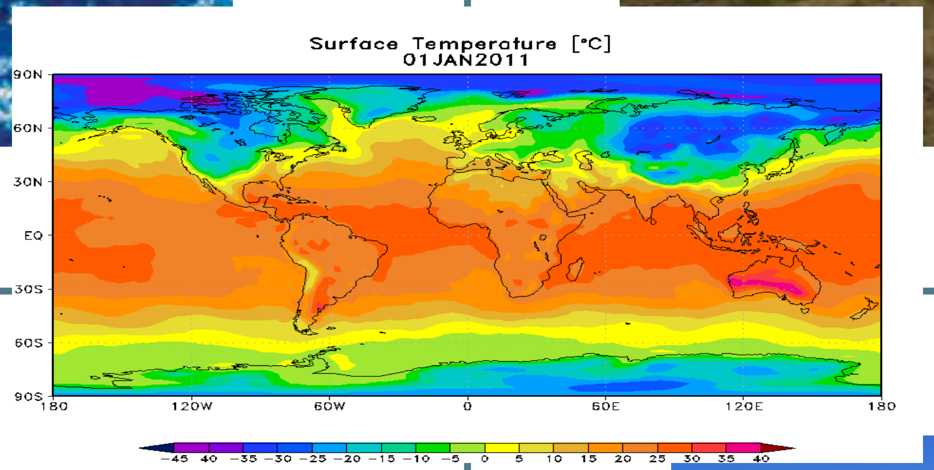
Regional hydrology exhibits large variations among major IPCC model projections

Low uncertainty	High uncertainty
Temperature	Hurricanes
Pressure	Extremes
Large-scale wind	Precipitation







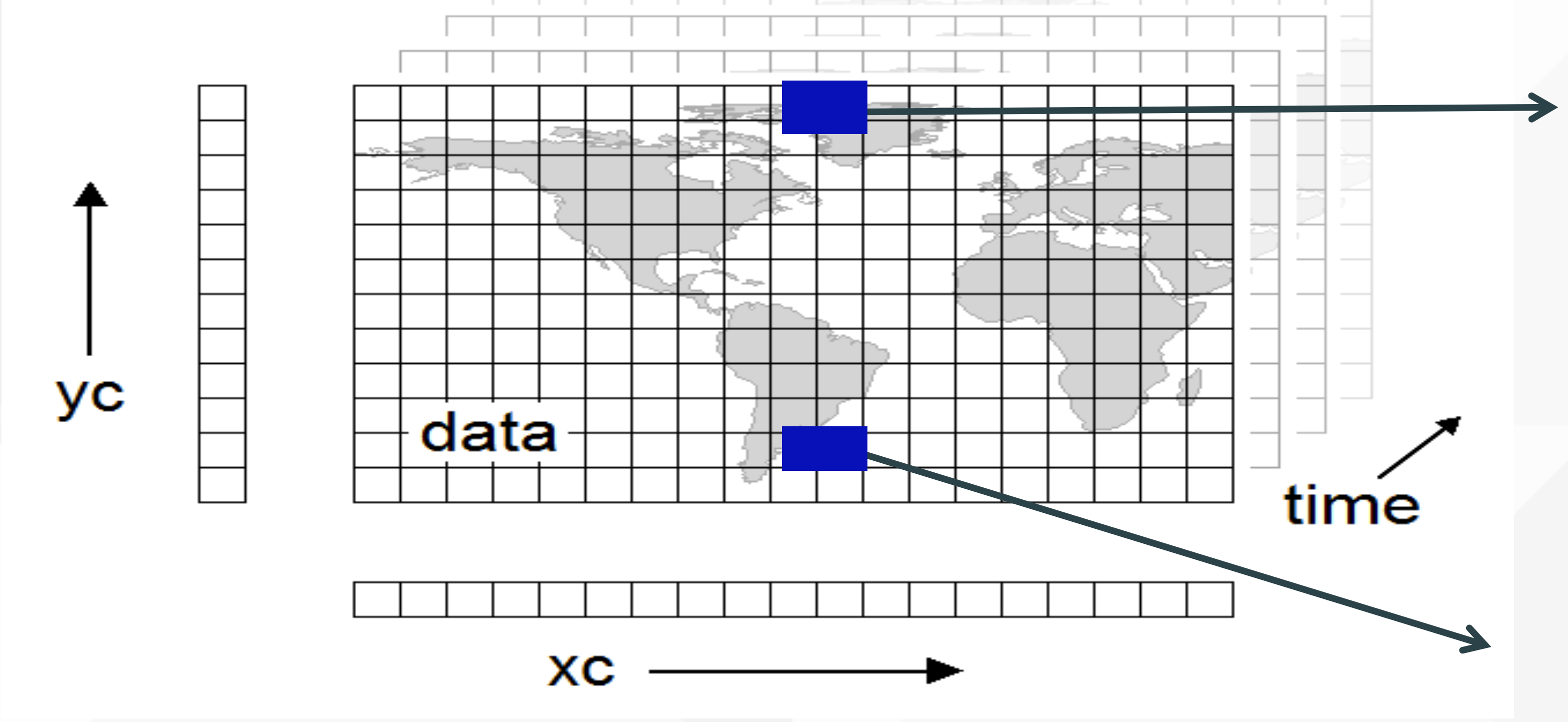


AI + HPC + Observation in Climate Science

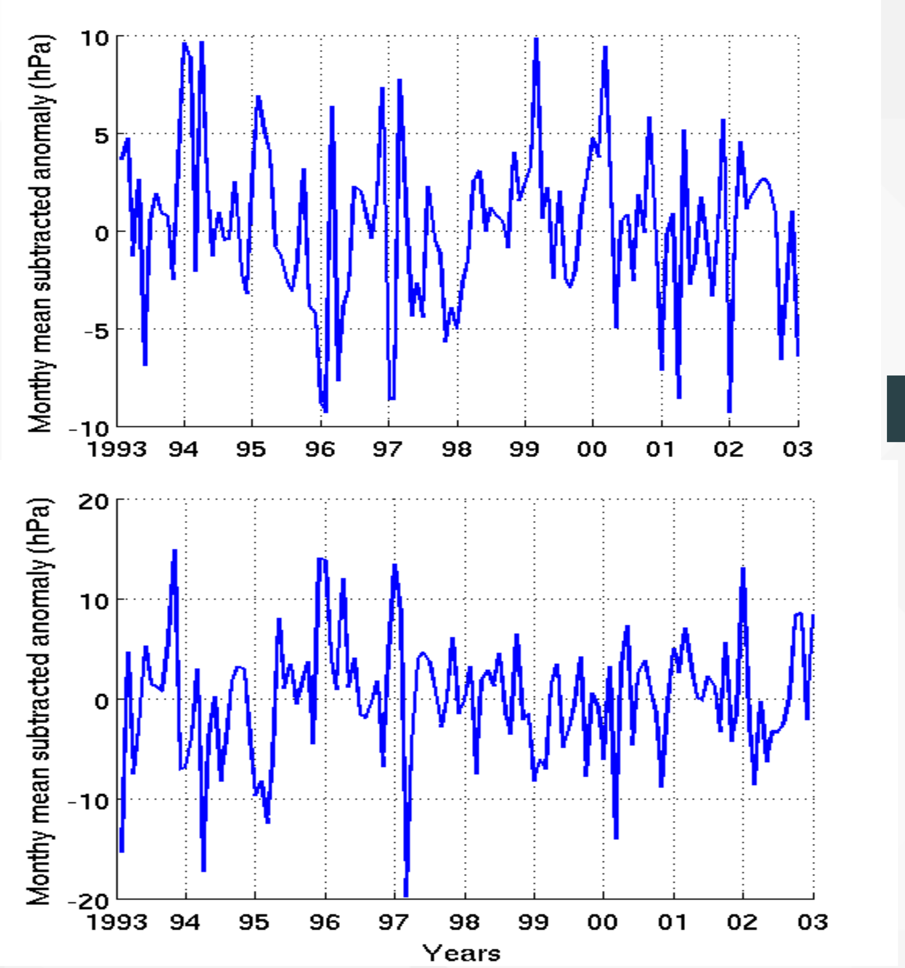
- Transformation from Data-Poor to Data-Rich : Make use of wealth of observational and simulation data
  - Accelerate Climate Models (PDE/ML)
  - Integrate Sensor Observations with Climate models (cloud/precipitation, land cover/biogeochem, sea ice/calibration)
  - AI/ML - Automated Model Extraction

AI/ML Driven Approach Illustration – Predicting Extreme Events

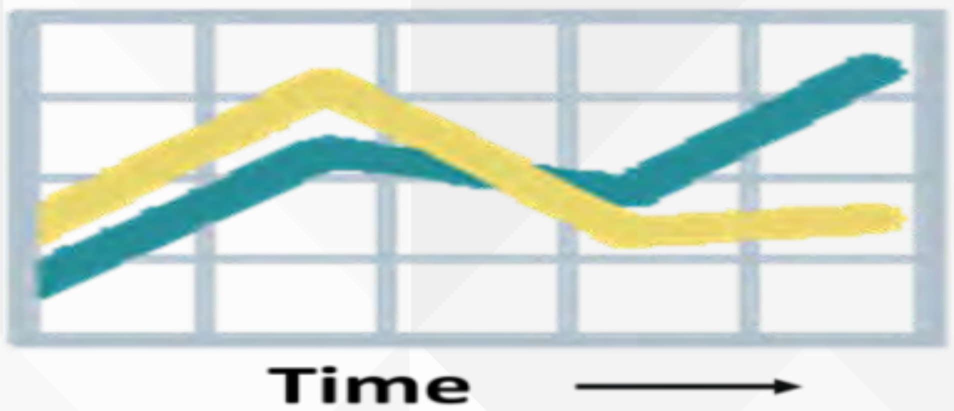
Climate Data



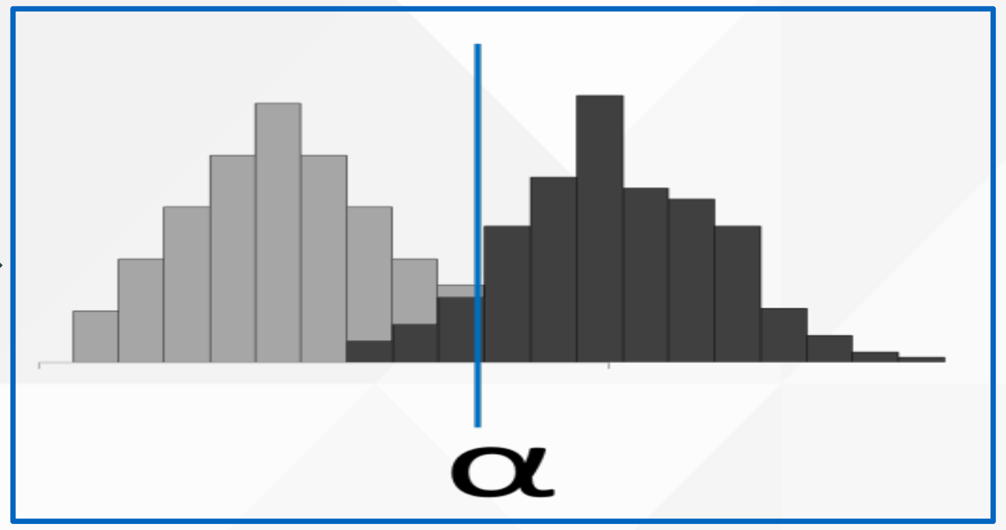
Simulation + Observation



Anomaly time series

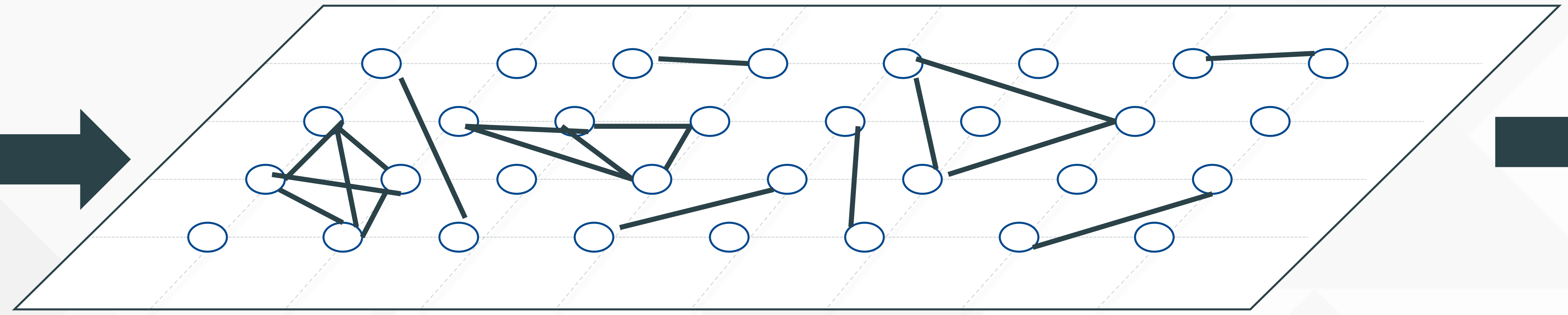


Correlation between anomaly time series/AR

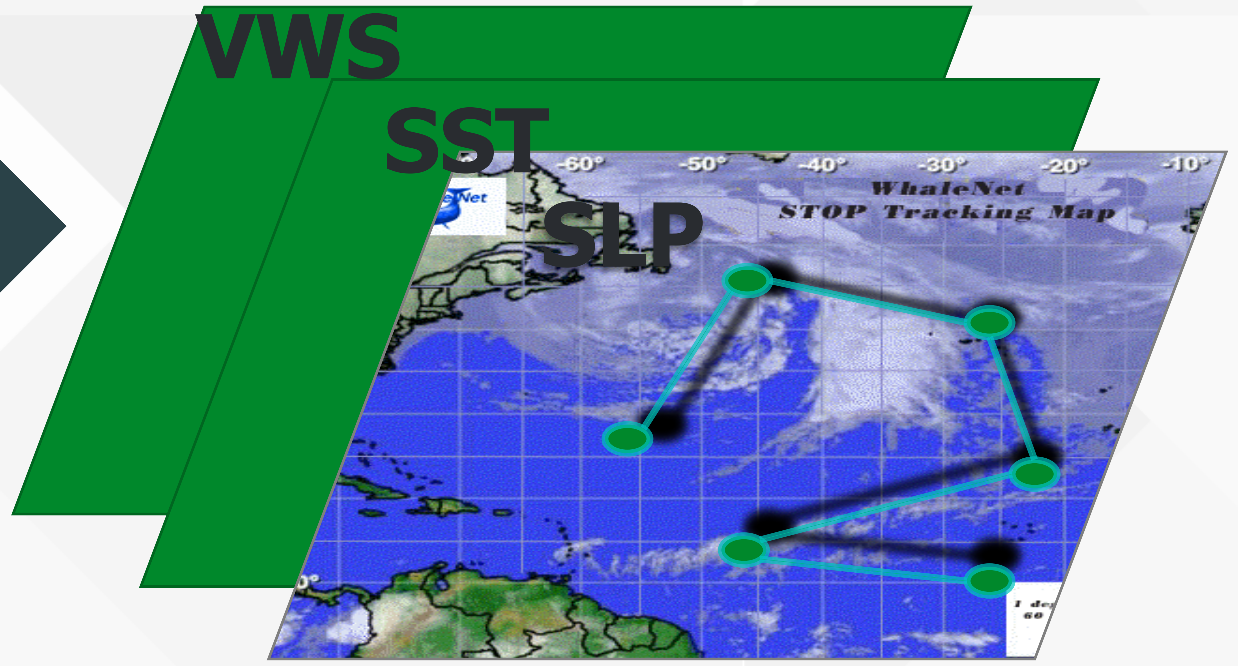


Stat. significant correlations

Climate Network



Edge weights: significant correlations
Nodes in the graph: grid points on the globe

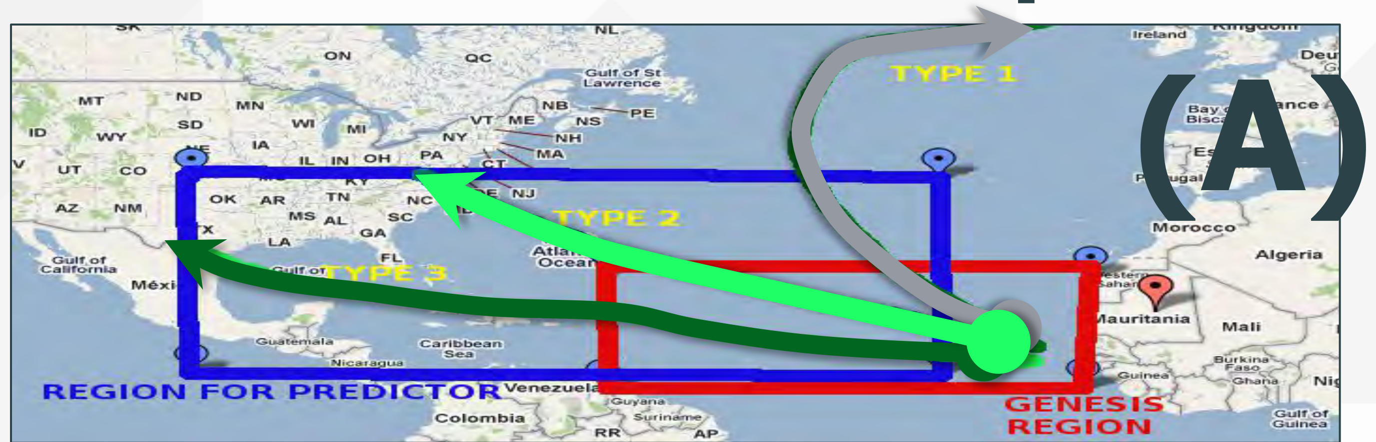


Multivariate Networks

Extreme Phase

Normal Phase

Multiphase Networks



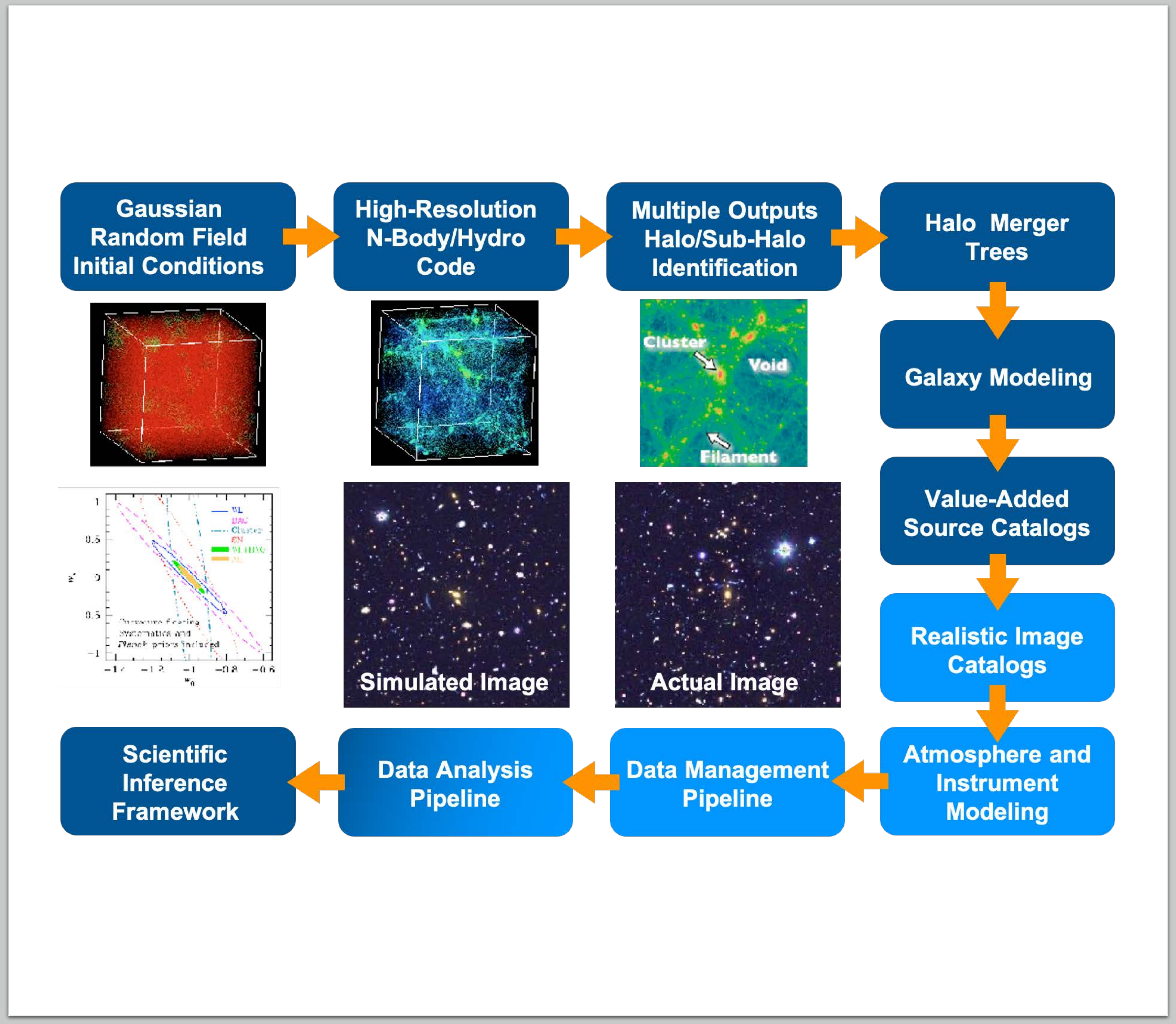
Micro, Real-time Forecast – Observation + AI/ML + Simulation

Cosmology*

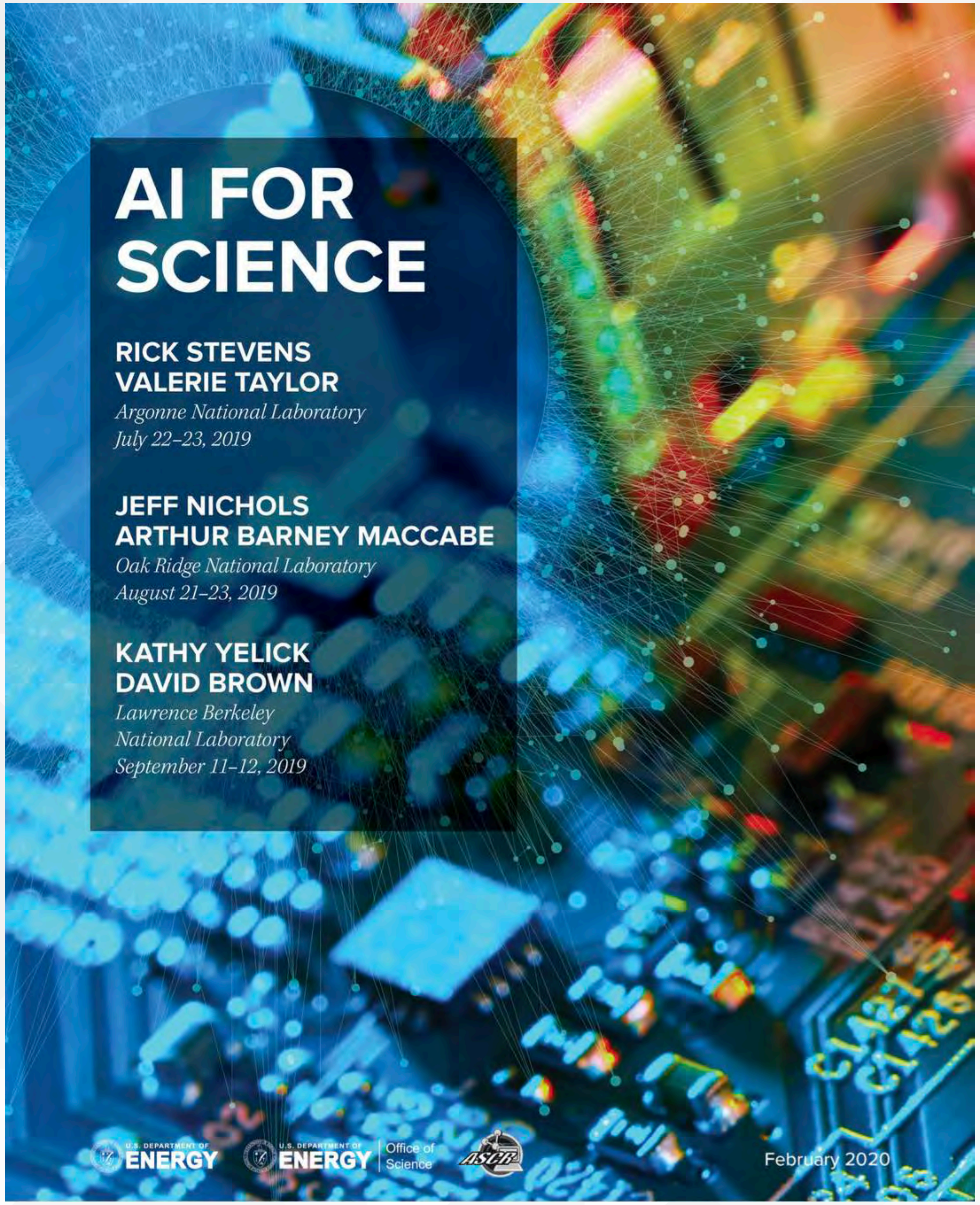
- Cosmic Frontier – AI in end-to-end application:
    - Precision Cosmic Microwave Background emulation – AI simulation speed-up of a factor of 1000
    - Search for strong lensing of galactic sources for precision cosmology measurements using AI classification, regression, and GANs for image generations

** [Argonne-led SciDAC-4 project: "Inference and Machine Learning at Extreme Scales"]*

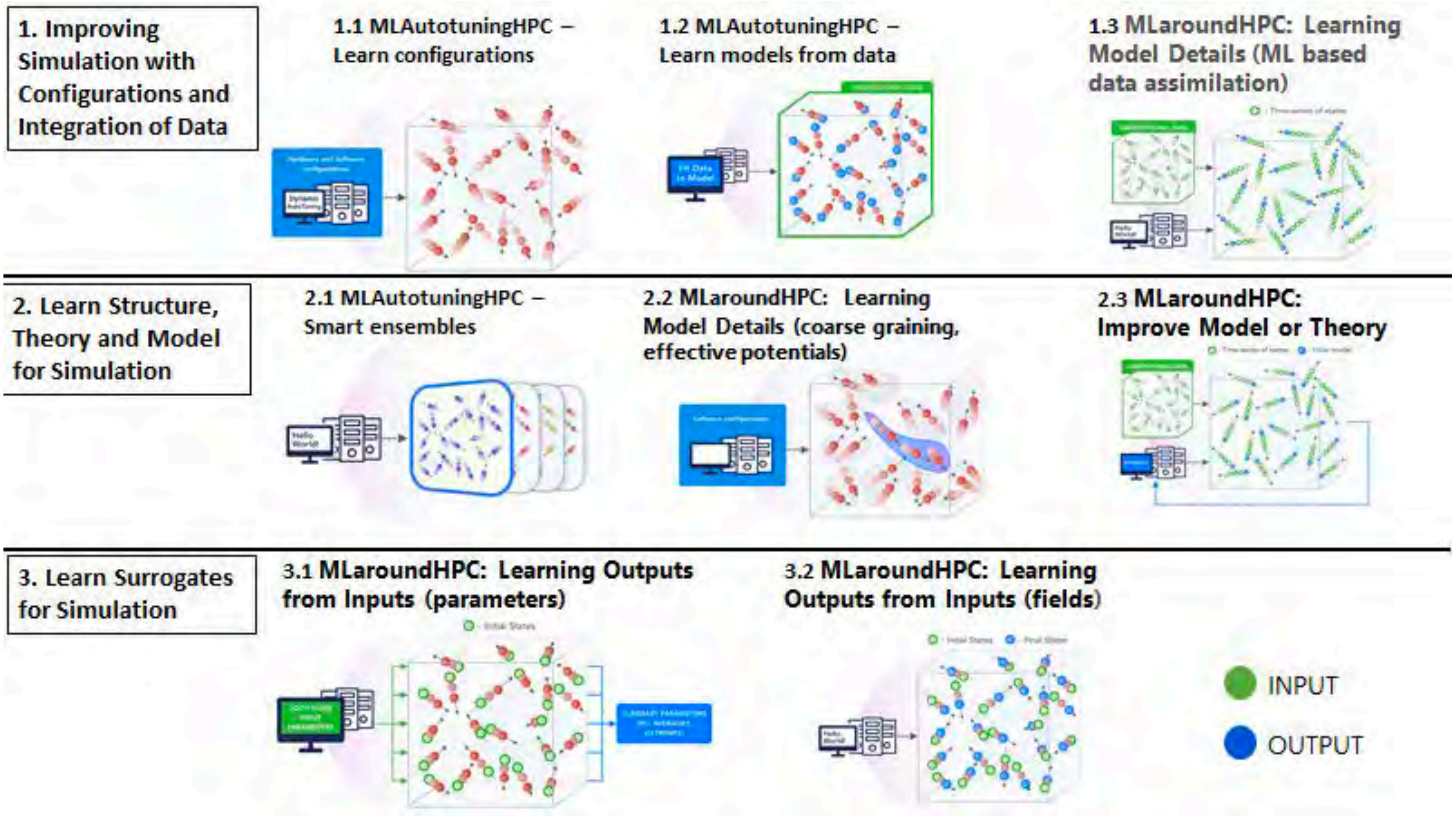
1) GANs for image emulation, 2) GP and DL-based emulators for summary statistics, 3) CNN-based image classification, 4) AI-based photometric reshift estimation, 5) Likelihood-free methods for inference



A Good Reference for Many Applications and CS



- DOE Organized - Over 1,000 scientists participated in town halls during the summer/fall of 2019
- Research Opportunities in AI
  - Biology, Chemistry, Materials,
  - Climate, Physics, Energy, Cosmology
  - Mathematics and Foundations
  - Data Life Cycle
  - Software Infrastructure
  - Hardware for AI
  - Integration with Scientific Facilities
  - https://www.anl.gov/ai-for-science-report



Geoffrey Fox, Shantenu Jha. "**Learning Everywhere: A Taxonomy for the Integration of Machine Learning and Simulations**" https://arxiv.org/abs/1909.13340

Next Set of Doors?

- Simulation in ML methods
- Surrogate AI functions
- Generative models to compare with simulation
- Learned functions
- Learned theory from data
- Guided search through parameter spaces
- Automated (no human in loop) complex workflow across paradigms of discovery
- … 1000X speed up

AI for Science

THANK YOU!

Dr. Alok Choudhary
Henry and Isabel Dever Professor
EECS and Kellogg School of Management
Northwestern University
choudhar@eecs.northwestern.edu

Founder, Chairman and Chief Scientist
4Cinsights Inc: A Big Data Science Company
(recently acquired by Mediaocean)
+1 312 515 2562
www.4cinsights.com

Lots of TOOLS at http://info.eecs.northwestern.edu

http://cucis.ece.northwestern.edu/publications/