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

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

Experiments

3rd paradigm: Computational science (simulations)

Density Functional Theory, Molecular Dynamics

4th paradigm: (Big) data driven science

Predictive analytics
Clustering
Relationship mining
Anomaly detection

Change in heat - Work done

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

Challenges
- Big Data
- Big Compute
- Architecture Search
- Interpretability

Advantages
- Feature Engineering Free
- More Accurate Models
- Faster Models

Performance

Amount of data

Deep learning
Most learning algorithms
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
Navier-Stokes Equations

Continuity Equation
\[ \nabla \cdot \mathbf{V} = 0 \]

Momentum Equations
\[ \rho \frac{D\mathbf{V}}{Dt} = -\nabla p + \rho \mathbf{g} + \mu \nabla^2 \mathbf{V} \]

\[ \frac{dx}{dt} = v \]
\[ \frac{dv}{dt} = \frac{F_{ext}}{m} \]
\[ x_2 - x_1 = v \Delta t \]
\[ v_2 = v_1 + \frac{F_{ext}}{m} \Delta t \]

Theory/Model Driven Point Workflow

- **Theory/Model**
- **Simulations**
- **Experiments/Observations**
Navier-Stokes Equations

Continuity Equation
\[ \nabla \cdot \vec{V} = 0 \]

Momentum Equations
\[ \rho \frac{D\vec{V}}{Dt} = -\nabla p + \rho \vec{g} + \mu \nabla^2 \vec{V} \]

Total derivative \( \partial \)
Pressure gradient \( \nabla p \)
Body force term \( \rho \vec{g} \)
Diffusion \( \mu \nabla^2 \vec{V} \)

Theory/Model Driven Point Workflow => Point Solution

- Theory/Model
- Simulations
- Experiments/Observations
Experiment/Observation
Driven Point Workflow

\[
\frac{dx}{dt} = v \\
\frac{dv}{dt} = \frac{F_{net}}{m} \\
x_2 - x_1 = vdt \\
v_2 = v_1 + \int_{t_1}^{t_2} \frac{F_{net}}{m} \, dt
\]
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

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

Engineering relationships of goals and means

Science relationships of cause and effect
Single AI/ML step applicable to multiple design problems

**Data**
- Hundreds of thousands of DFT calculations (e.g., OQMD)

**Composition-based models**
- 145 attributes (stoichiometric/elemental/electronic/ionic)
  - Mean Electronegativity
  - Bond Ionic Character... (148+ properties)

**Structure-aware models**
- Voronoi tessellations to capture local environment of atoms

**Inverse models**
- Stable compounds, metallic glasses, semiconductors?

Online Tool: [http://info.eecs.northwestern.edu/FEpredictor](http://info.eecs.northwestern.edu/FEpredictor)
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

- Experiments
- Simulations
Semiconductors – E.g., Formation Energy

- Fingerprint of entire unexplored ternary composition space!

- **Inference Engine**

**Measured Glass Formation**

**Predicted Glass Formation**

<table>
<thead>
<tr>
<th>Composition</th>
<th>$E_g$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScHg$_4$Cl$_7$</td>
<td>1.26</td>
</tr>
<tr>
<td>V$_2$Hg$_3$Cl$_7$</td>
<td>1.16</td>
</tr>
<tr>
<td>Mn$_6$CCl$_8$</td>
<td>1.28</td>
</tr>
<tr>
<td>Hf$_5$Si$_1$Cl$_2$</td>
<td>1.11</td>
</tr>
<tr>
<td>VCu$_5$Cl$_9$</td>
<td>1.19</td>
</tr>
</tbody>
</table>

New Semiconductors

New metallic glasses

Alloy composition

- $\text{Zr}_{0.38}\text{Co}_{0.24}\text{Cu}_{0.38}$
- $\text{V}_{0.16}\text{Ni}_{0.64}\text{B}_{0.2}$
- $\text{Zr}_{0.46}\text{Cr}_{0.36}\text{Ni}_{0.18}$
- $\text{Zr}_{0.5}\text{Fe}_{0.38}\text{W}_{0.12}$
**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.

---

**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
- $\text{MAE}_{FE} = \sim 0.06$ eV/atom

---

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Training from Scratch</th>
<th>Transfer Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAE (eV/atom)</td>
<td>MAE (eV/atom)</td>
</tr>
<tr>
<td>OQMD</td>
<td>341,000</td>
<td>0.0437</td>
<td>-</td>
</tr>
<tr>
<td>JARVIS</td>
<td>11,050</td>
<td>0.0568</td>
<td>0.0312</td>
</tr>
<tr>
<td>Materials Project</td>
<td>23,641</td>
<td>0.0327</td>
<td>0.0247</td>
</tr>
<tr>
<td>Experimental</td>
<td>1,963</td>
<td>0.1460</td>
<td>0.0608</td>
</tr>
</tbody>
</table>

Jha et al., *Nature Communications*, 2019
Example: Industrial Materials Design
A complex and expensive work-flow

- Prediction of properties of samples under certain processing conditions
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

Property Prediction – numerical data

<table>
<thead>
<tr>
<th>Folder</th>
<th>File name</th>
<th>Machine</th>
<th>Mode</th>
<th>Target</th>
<th>position (fp)</th>
<th>magnification</th>
</tr>
</thead>
<tbody>
<tr>
<td>T180223</td>
<td>01/01/02</td>
<td>SEM</td>
<td>COMPO</td>
<td>powder</td>
<td>C00</td>
<td>x200</td>
</tr>
<tr>
<td>B180831</td>
<td>01/01/02</td>
<td>SEM</td>
<td>COMPO</td>
<td>forge</td>
<td>C00</td>
<td>x30000</td>
</tr>
<tr>
<td></td>
<td>01/01/03</td>
<td>SEM</td>
<td>COMPO</td>
<td>powder</td>
<td>C00</td>
<td>x200</td>
</tr>
<tr>
<td></td>
<td>01/01/04</td>
<td>SEM</td>
<td>COMPO</td>
<td>forge</td>
<td>C00</td>
<td>x30000</td>
</tr>
<tr>
<td></td>
<td>01/01/05</td>
<td>SEM</td>
<td>COMPO</td>
<td>powder</td>
<td>C00</td>
<td>x200</td>
</tr>
<tr>
<td></td>
<td>01/01/06</td>
<td>SEM</td>
<td>COMPO</td>
<td>forge</td>
<td>C00</td>
<td>x30000</td>
</tr>
<tr>
<td></td>
<td>01/01/07</td>
<td>SEM</td>
<td>COMPO</td>
<td>powder</td>
<td>C00</td>
<td>x200</td>
</tr>
<tr>
<td></td>
<td>01/01/08</td>
<td>SEM</td>
<td>COMPO</td>
<td>forge</td>
<td>C00</td>
<td>x30000</td>
</tr>
</tbody>
</table>
Deep (transfer) learning*

* What DL enables - Not enough Experimental data to learn
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)
  o Band gap and glass forming ability prediction [npjCM 2016]
  o Bulk modulus prediction [RSC Adv 2016]
  o Seebeck coefficient prediction [JCompChem 2018]
  o Chemical properties prediction [NIPS MLMM 2018, IJCNN 2019, Molecular Informatics 2019]

• Inverse PSPP models (optimization/discovery)
  o Stable compounds [PRB 2014]
  o Magnetostrictive materials [Nature Scientific Reports 2015, AIAA 2018]
  o Semiconductors and metallic glasses [npjCM 2016]
  o Microstructure design (GAN) [JMD 2018]
  o Titanium aircraft panels [CMS 2019]

• Structure characterization
  o EBSD Indexing [BigData-ASH 2016, M&M 2018]
  o Crack detection in macroscale images [CBM 2017, IJTTE 2018]
  o XRD analysis for phase detection [IJCNN 2019]
  o Plastic deformation identification [IJCNN 2019]

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)
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**
- Data
- Time
- Climate Network:
  - Nodes: Grid points on the globe
  - Edge weights: Significant correlations

**Simulation + Observation**
- Anomaly time series

**Correlation between anomaly time series/AR**

**Temperature**

**Stat. significant correlations**

**Climate Data**
- SLP
- SST
- VWS

**Multivariate Networks**
- Extreme Phase
- Normal Phase

**Multiphase Networks**

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

Observe, Learn and Improve

HPC+AI/ML

High-resolution simulation

Predict and Act

High-resolution observation

Courtsey: RIKEN (JAPAN) - collaboration
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
1. Improving Simulation with Configurations and Integration of Data
   1.1 MLAutotuningHPC – Learn configurations
   1.2 MLAutotuningHPC – Learn models from data
   1.3 MLaroundHPC: Learning Model Details (ML based data assimilation)

2. Learn Structure, Theory and Model for Simulation
   2.1 MLAutotuningHPC – Smart ensembles
   2.2 MLaroundHPC: Learning Model Details (coarse graining, effective potentials)
   2.3 MLaroundHPC: Improve Model or Theory

3. Learn Surrogates for Simulation
   3.1 MLaroundHPC: Learning Outputs from Inputs (parameters)
   3.2 MLaroundHPC: Learning Outputs from Inputs (fields)

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

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

THANK YOU!

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

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