Lightning Overview of Machine Learning

Shinjae Yoo
Computational Science Initiative

70 YEARS OF DISCOVERY

A CENTURY OF SERVICE

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

Outline

- Why is Machine Learning important?
- Machine Learning Concepts
- Big Data and Machine Learning
- Potential Research Areas

Why is Machine Learning important?

ML Application to Physics

Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning

Matthias Rupp,[1,2] Alexandre Tkatchenko,[3,2] Klaus-Robert Müller,[1,2] and O. Anatole von Lilienfeld[4,2,*]

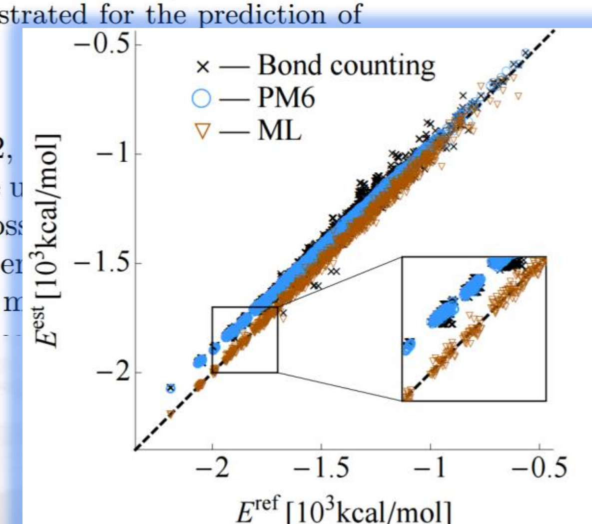
[1] *Machine Learning Group, Technical University of Berlin, Franklinstr 28/29, 10587 Berlin, Germany*
[2] *Institute of Pure and Applied Mathematics, University of California Los Angeles, Los Angeles, CA 90095, USA*
[3] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*
[4] *Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA*

(Dated: September 14, 2011)

We introduce a machine learning model to predict atomization energies of a diverse set of organic molecules, based on nuclear charges and atomic positions only. The problem of solving the molecular Schrödinger equation is mapped onto a non-linear statistical regression problem of reduced complexity. Regression models are trained on and compared to atomization energies computed with hybrid density-functional theory. Cross-validation over more than seven thousand small organic molecules yields a mean absolute error of ~10 kcal/mol. Applicability is demonstrated for the prediction of molecular atomization potential energy curves.

Solving the Schrödinger equation (SE), $H\Psi = E\Psi$, for assemblies of atoms is a fundamental problem in quantum mechanics. Alas, solutions that are exact up to numerical precision are intractable for all but the smallest systems with very few atoms. Hierarchies of approximations have

ory (DFT) level [2, of theory could be u ML training. Cross a mean absolute er der of magnitude m



ML Application to Biology

commentary

An active role for machine learning in dru

Robert F Murphy

Because of the complexity of biological systen
for future drug development. In particular, ma
imaging assays and active-learning methods t
dimensionality problem in drug development.

High-throughput and high-content screening have been widely adopted by pharmaceutical and biotechnology companies as well as by many academic labs over the past 20 years, with the goal of rapidly identifying potential drugs that affect specific molecular targets[1–3]. These technologies dramatically enhance the rate and amount of information that can be collected about the effects of chemical compounds, and publicly funded efforts such as the Molecular Libraries Screening Centers of the US National Institutes of Health have permitted the creation of

models, is w
machine lea
important r
and develop
Here I focus
learning can
use of machi
information
assays and t
learning to c

Seeing mo
High-throug
content scre

REVIEWS

Machine learning applications in genetics and genomics

Maxwell W. Libbrecht[1] and William Stafford Noble[1,2]

Abstract | The field of machine learning, which aims to develop computer algorithms that improve with experience, holds promise to enable computers to assist humans in the analysis of large, complex data sets. Here, we provide an overview of machine learning applications for the analysis of genome sequencing data sets, including the annotation of sequence elements and epigenetic, proteomic or metabolomic data. We present considerations and recurrent challenges in the application of supervised, semi-supervised and unsupervised machine learning methods, as well as of generative and discriminative modelling approaches. We provide general guidelines to assist in the selection of these machine learning methods and their practical application for the analysis of genetic and genomic data sets.

Machine learning

The field of machine learning is concerned with the

regulatory elements followed by sequencing (FAIRE-

U.S. DEPARTMENT OF ENERGY

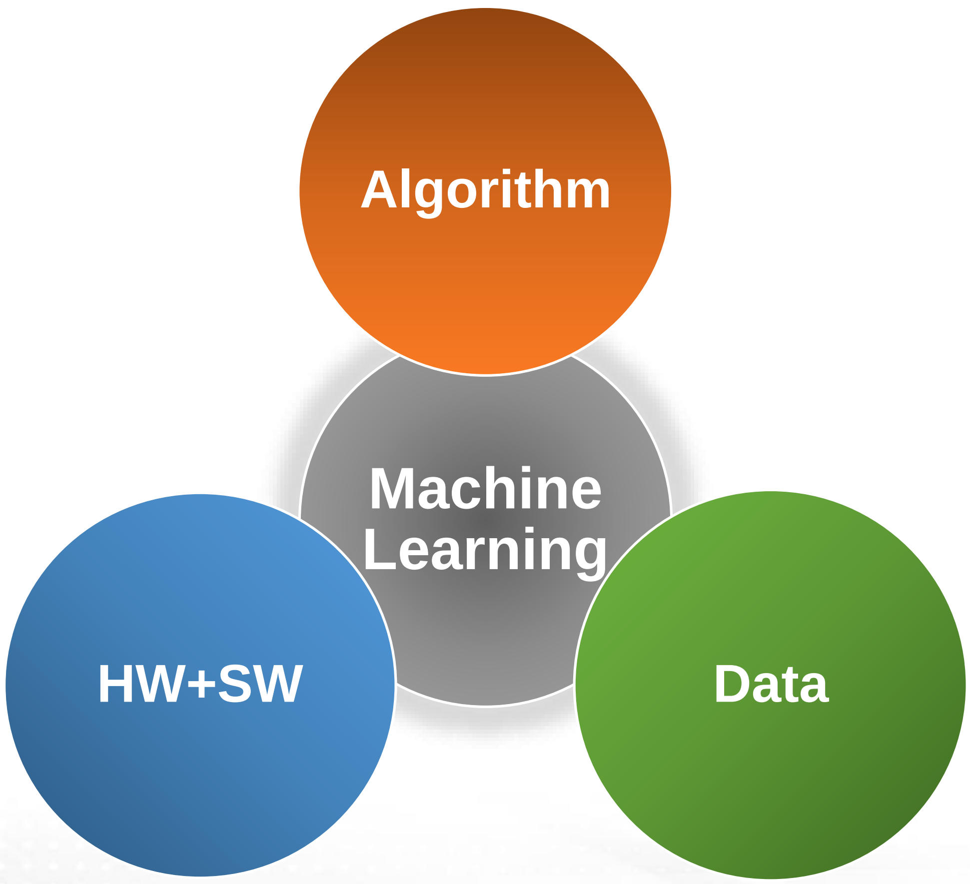
BROOKH
NATIONAL LAB

Machine Learning Concepts

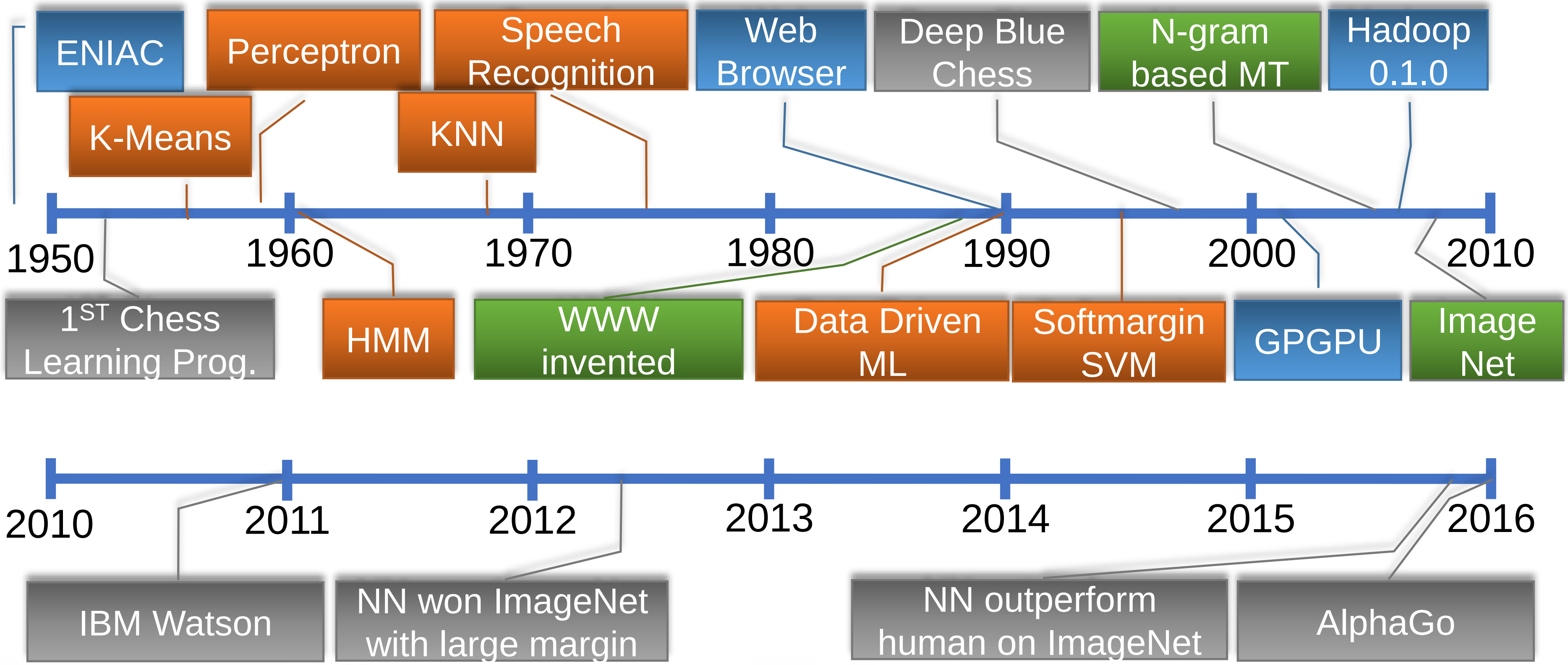
What is Machine Learning (ML)

- One of Machine Learning definitions
  - “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” Tom Mitchell, 2006
  - Statistics: What conclusions can be inferred from data
  - ML incorporates additionally
    - What architectures and algorithms can be used to effectively handle data
    - How multiple learning subtasks can be orchestrated in a larger system, and questions of computational tractability

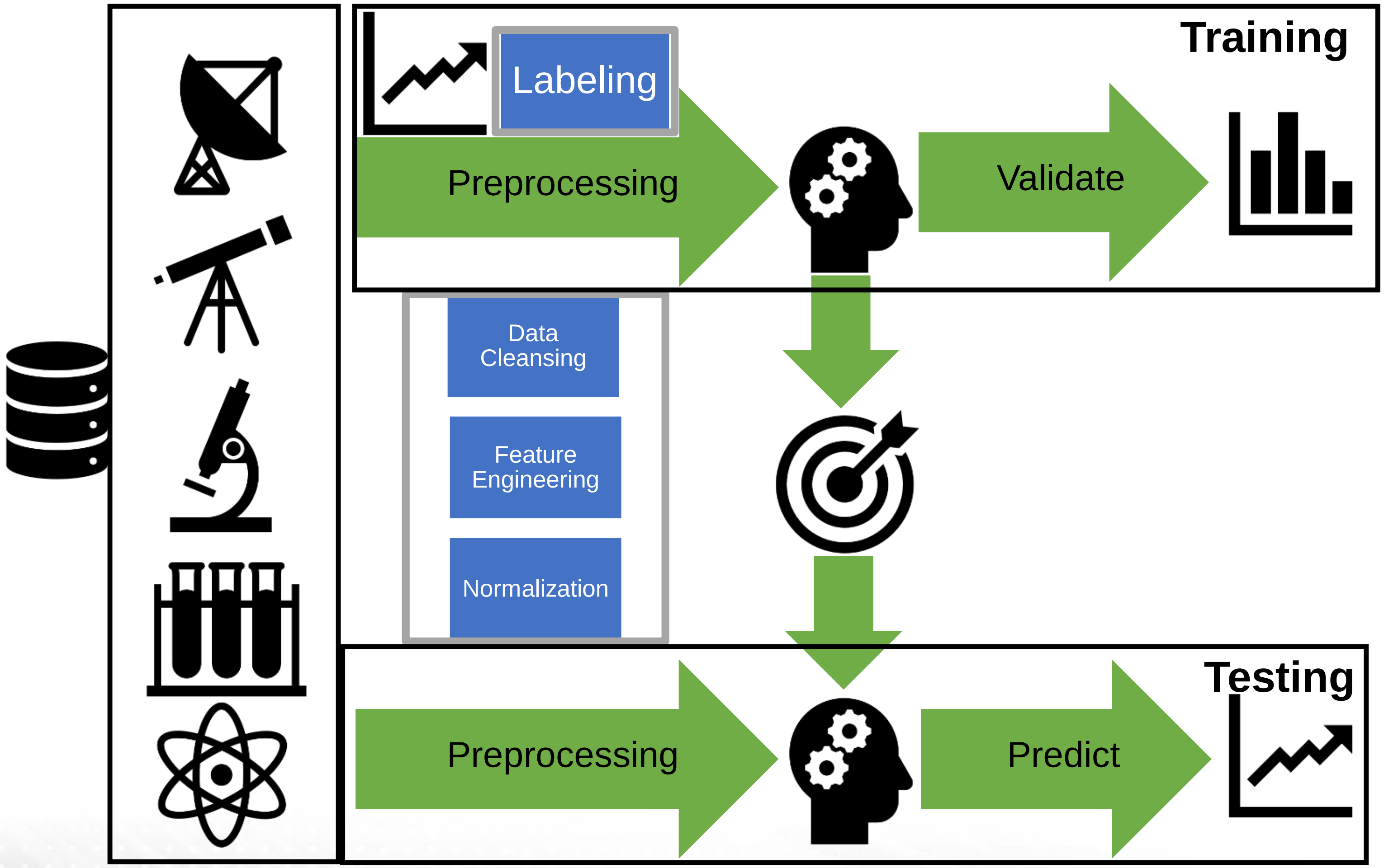
Machine Learning Components



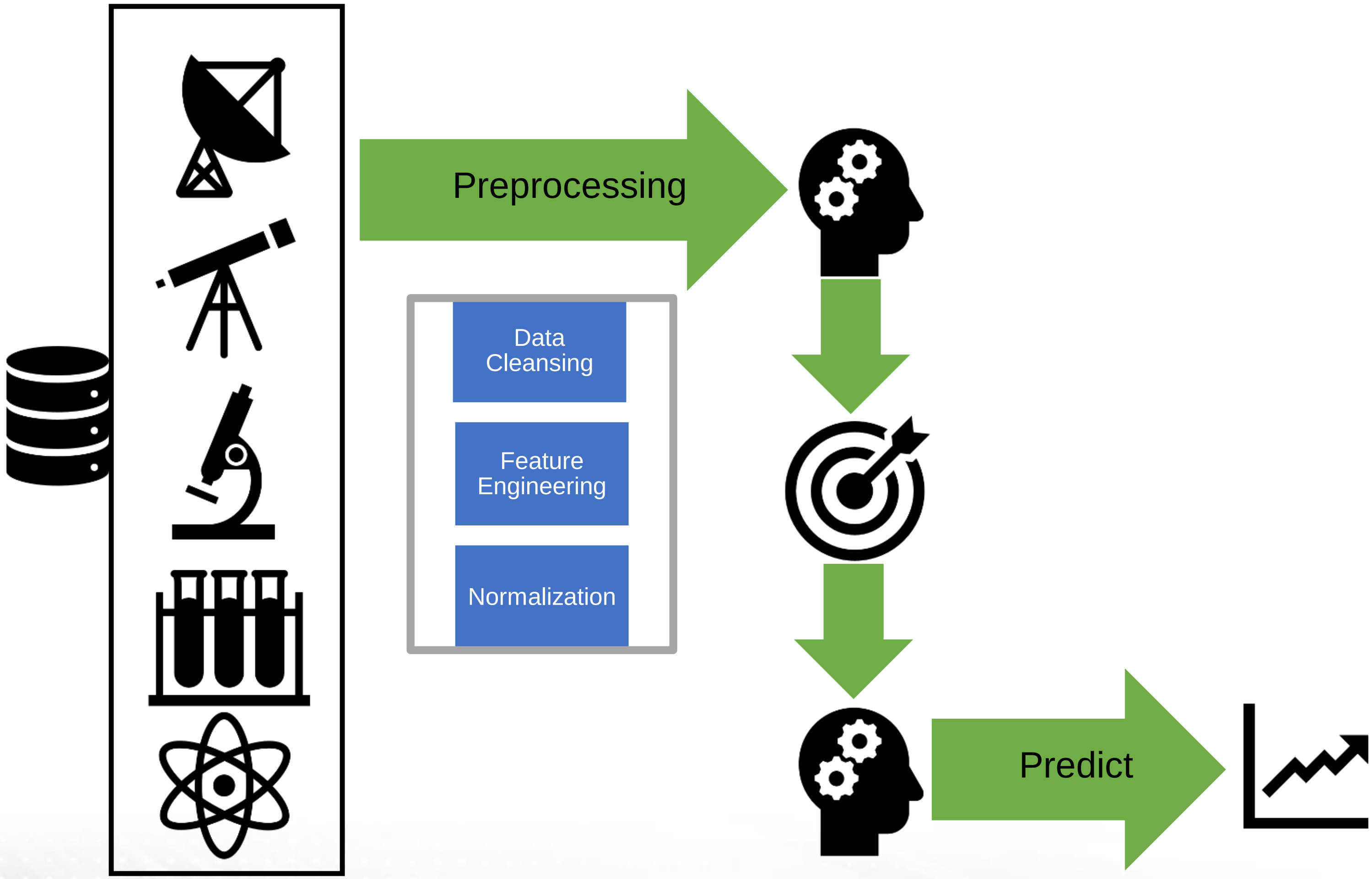
Brief History of Machine Learning



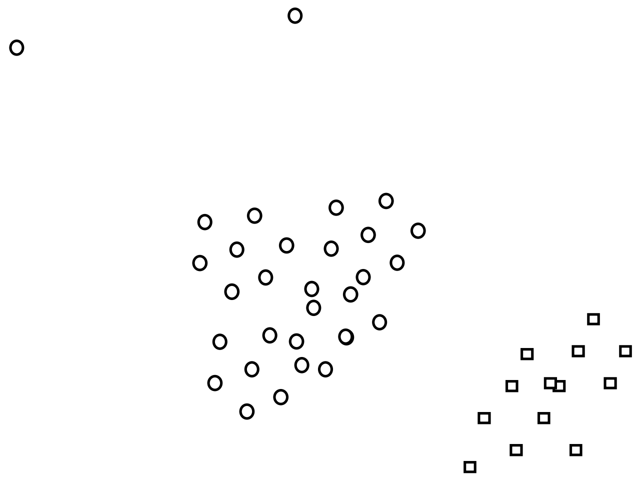
Supervised Learning Pipeline



Unsupervised Learning Pipeline

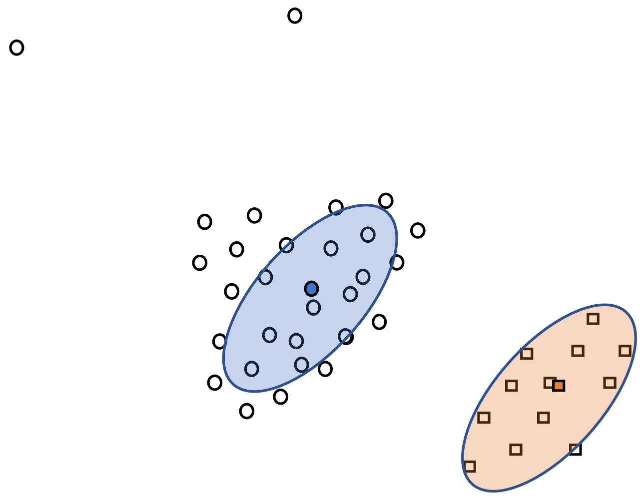


Types of Learning



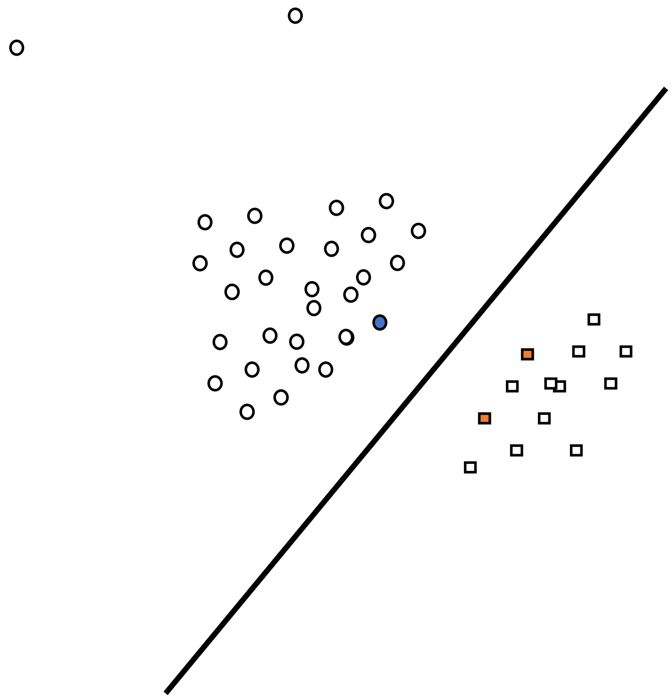
Types of Learning

- **Generative Learning**



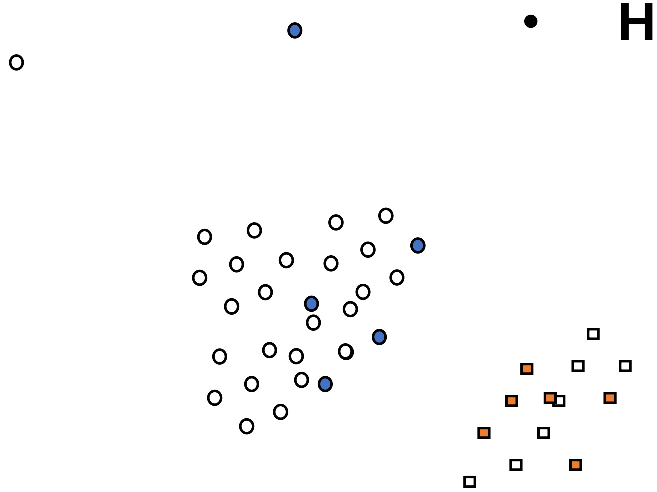
Types of Learning

- **Discriminative Learning**



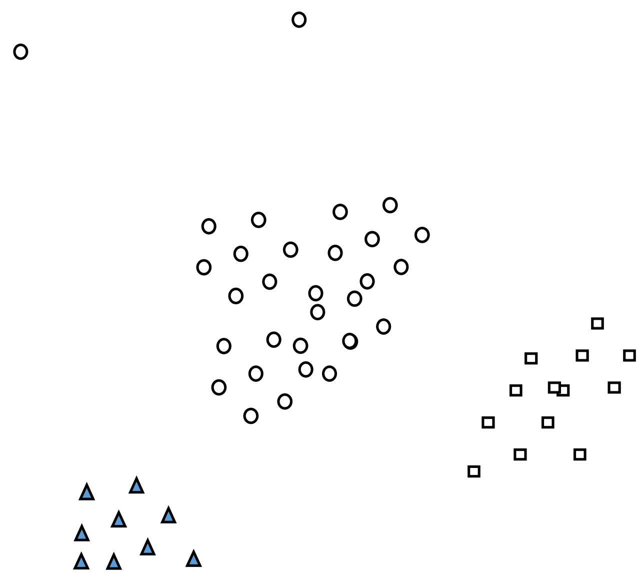
Types of Learning

- **Active Learning**
  - **How to select training data?**



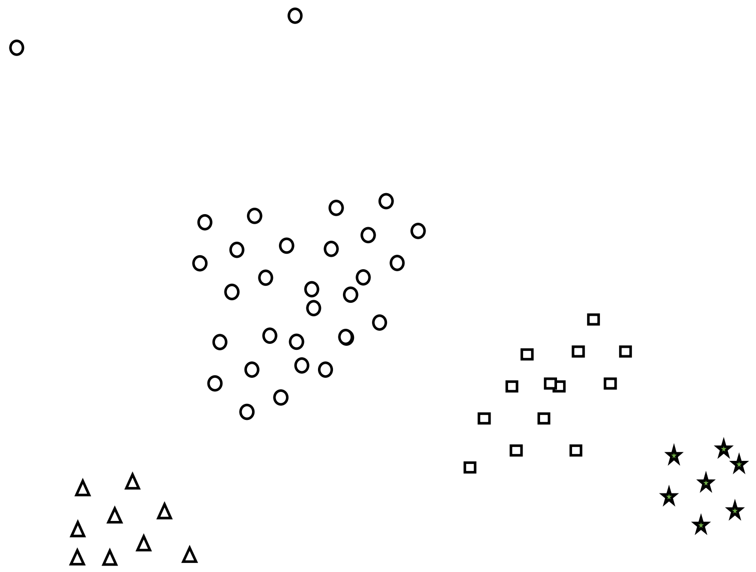
Types of Learning

- **Multi-task Learning**



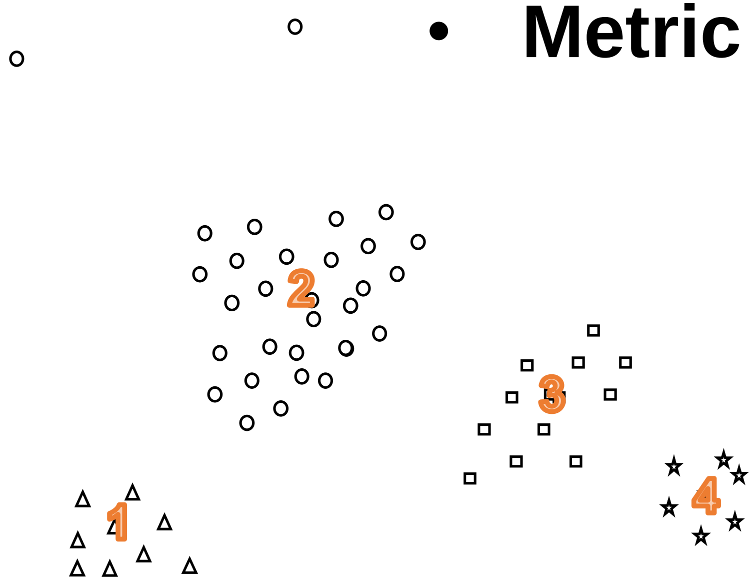
Types of Learning

- **Transfer Learning**



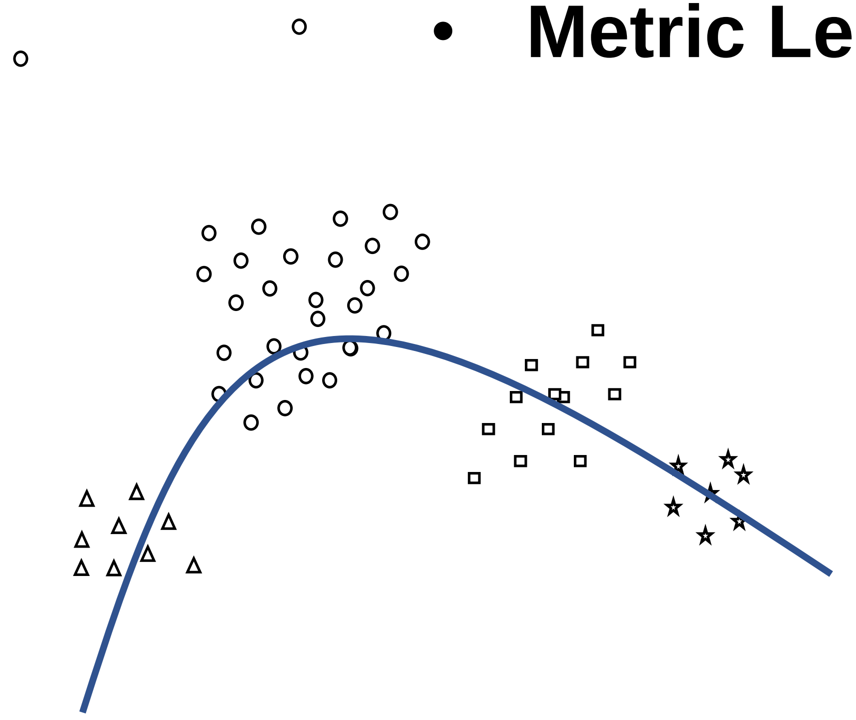
Types of Learning

- **Kernel Learning**
- **Metric Learning**



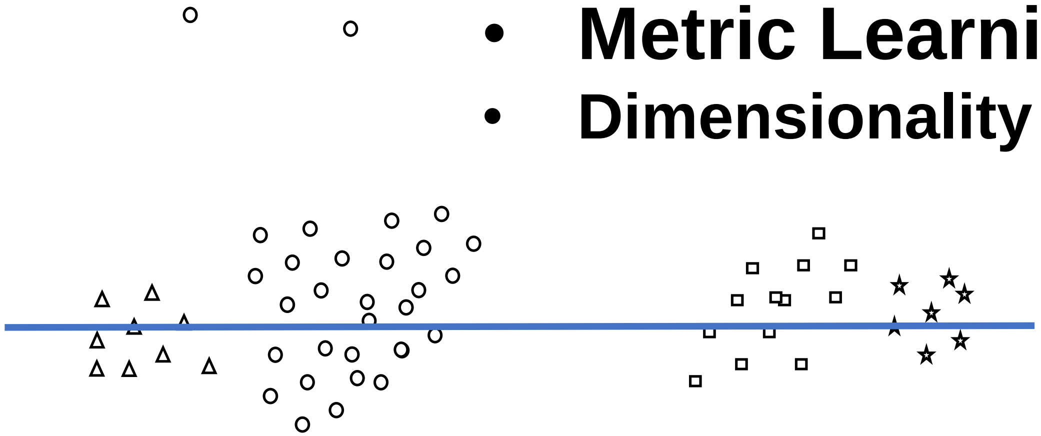
Types of Learning

- **Kernel Learning**
- **Metric Learning**



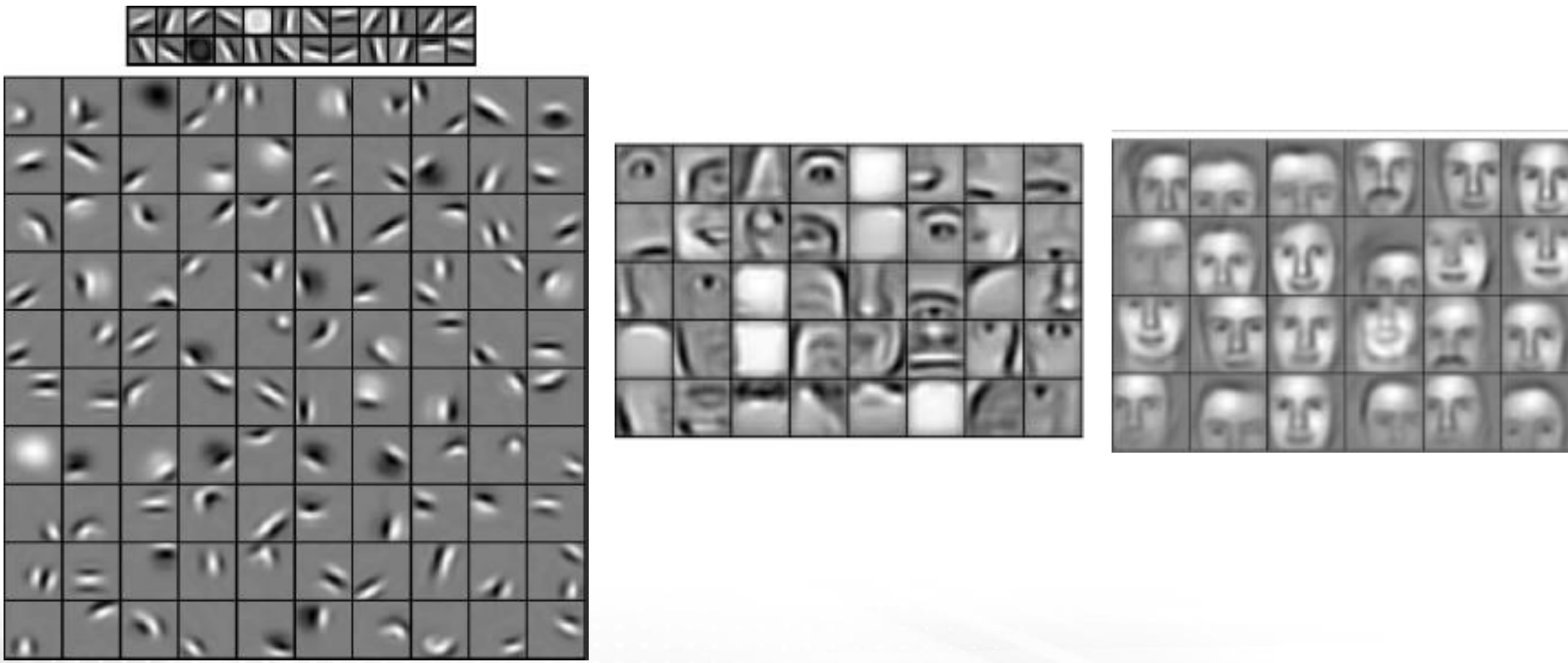
Types of Learning

- **Kernel Learning**
- **Metric Learning**
- **Dimensionality Reduction**



Types of Learning

- Feature Learning



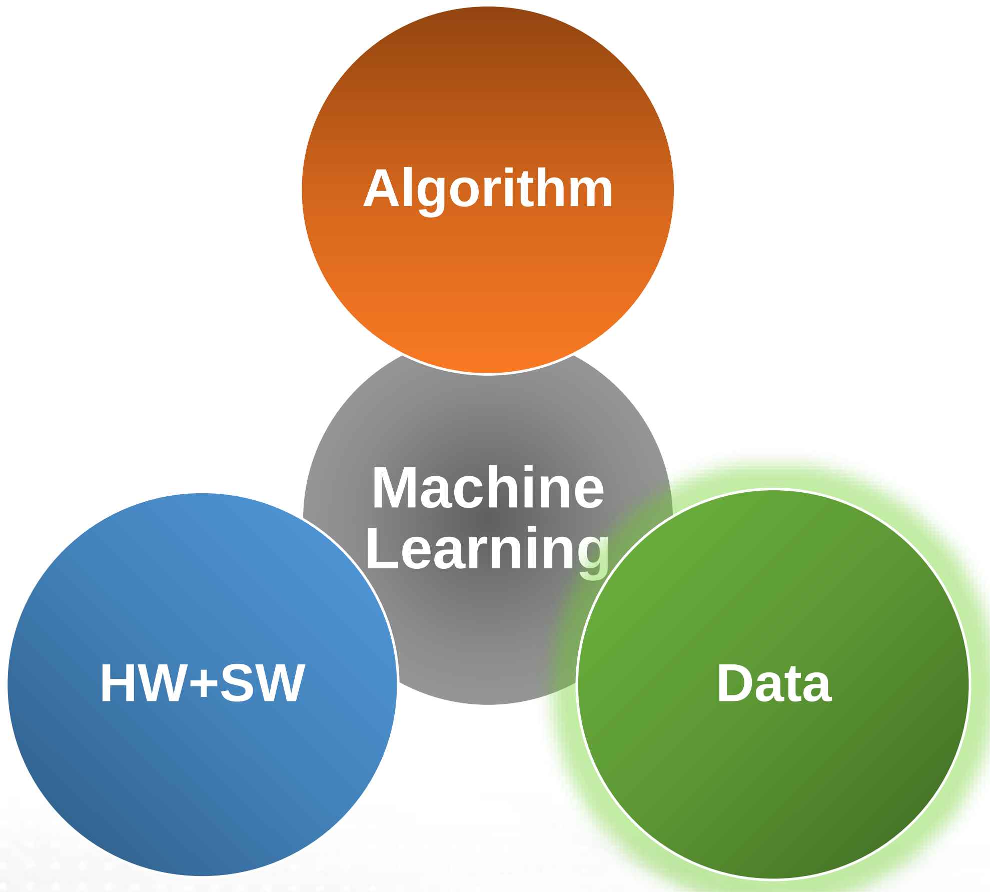
- Lee, et al. "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations", ICML '09

Machine Learning Algorithms

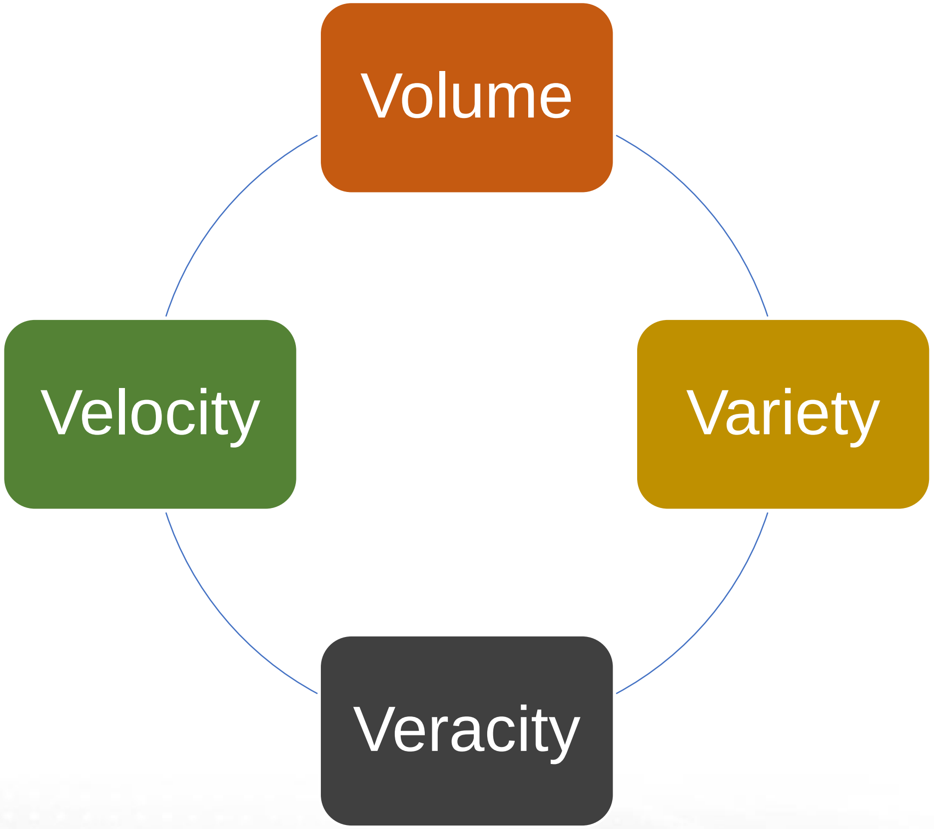
- Bayesian Algorithms
- Instance-based Algorithms
- Regularization Algorithms
- Decision Trees
- Association Rule Mining
- Ensemble Learning

Machine Learning with Big Scientific Data

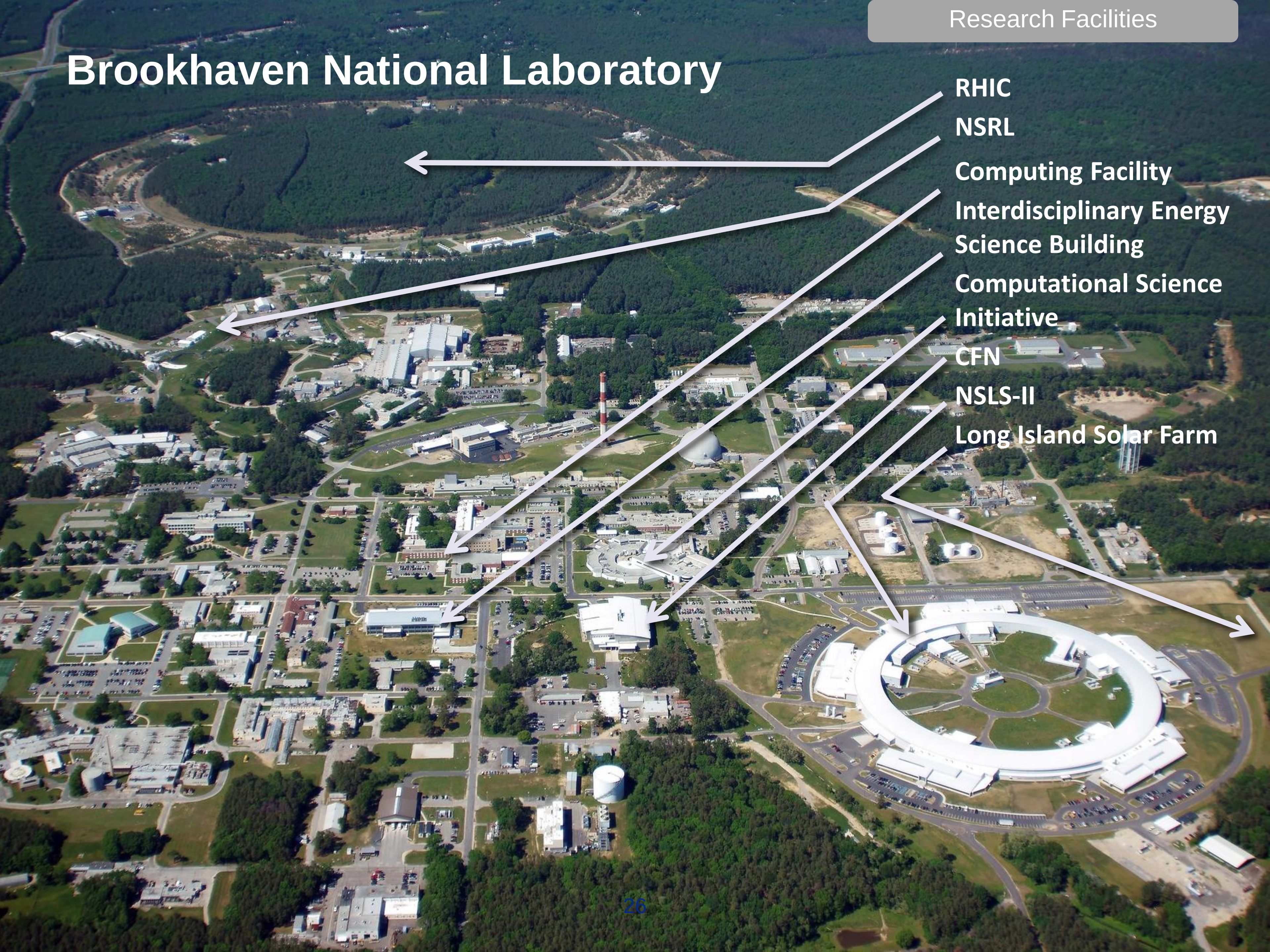
Machine Learning Components



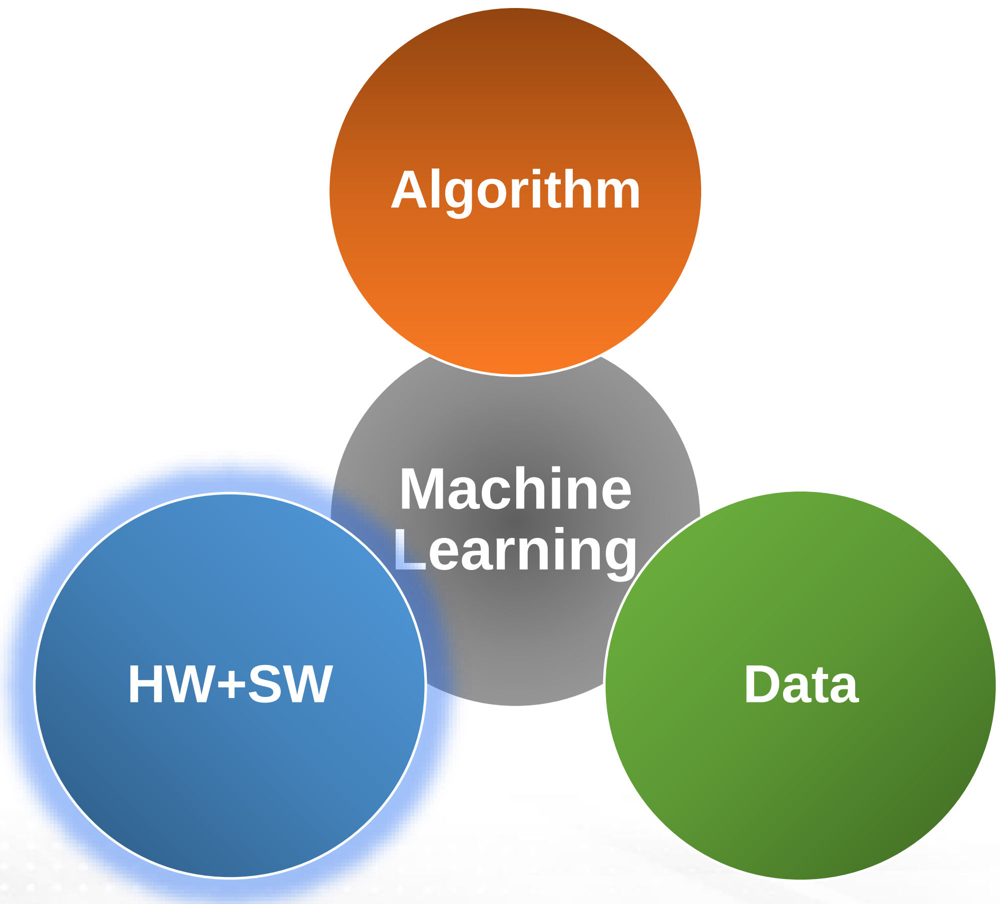
Big Data



Brookhaven National Laboratory

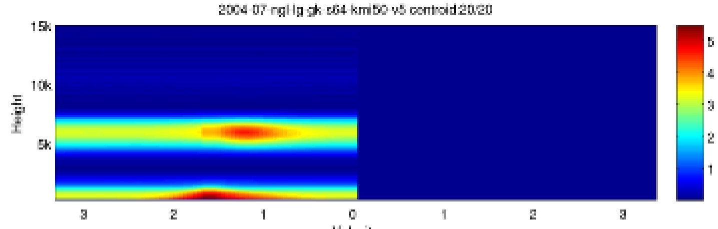


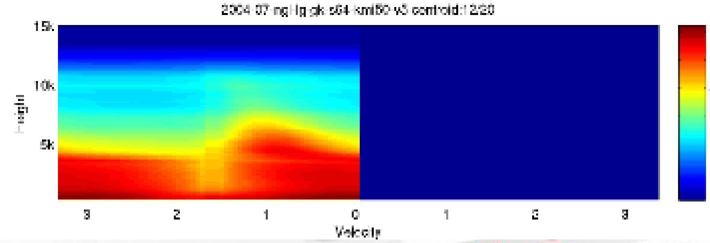
Machine Learning Components



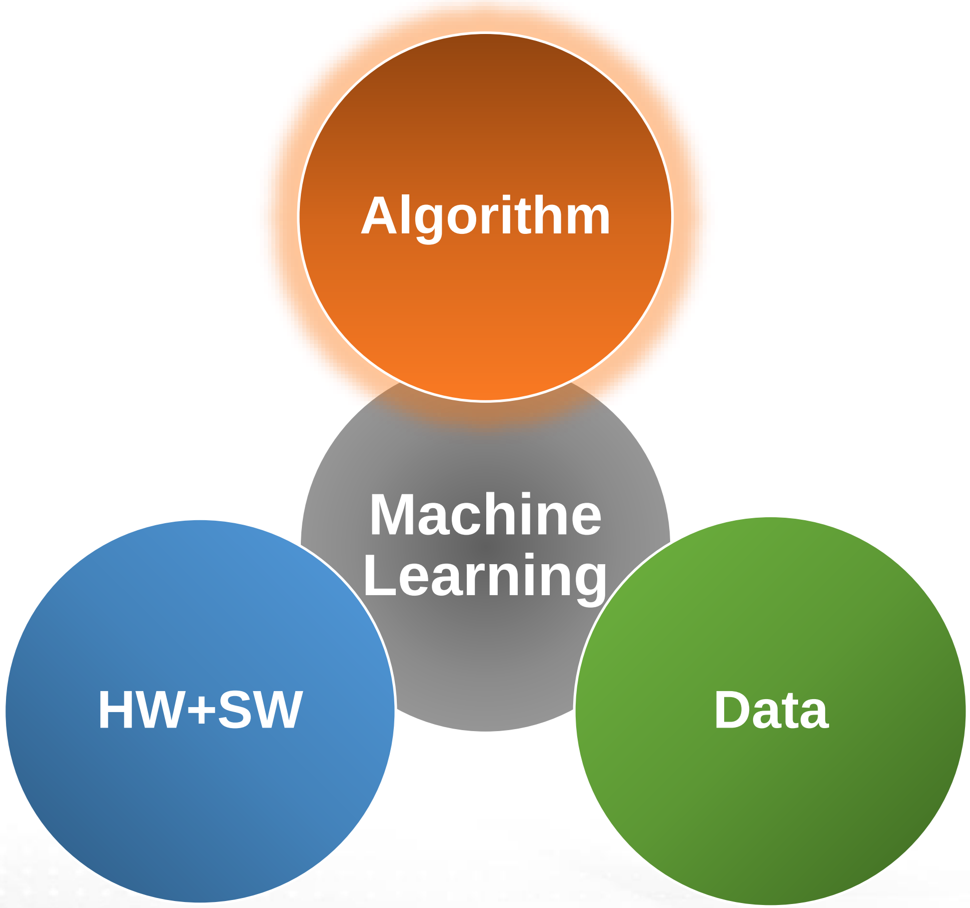
MapReduce: Not Complete Solution in 2010

- **Task**: Find cluster patterns in Doppler Radar Spectra
- **Data**: 1hr≈130MB, 1yr ≈1TB, 2004~2008 ≈ 5TB
- MapReduce (K-Means)
  - Map: Find closest centroids
  - Reduce: Update centroids
- MapReduce (Spectral Clustering)
  - Distributed Affinity Matrix Computation : O(n^2)
  - Distributed Lanczos Methods to compute EVD
- Scalability Analysis
  - 12 cores (1 node) Spectral clustering took 1 week for one month data
  - 616 cores (77 nodes) Spectral Clustering took less than 2 hours for three months (~300GB)

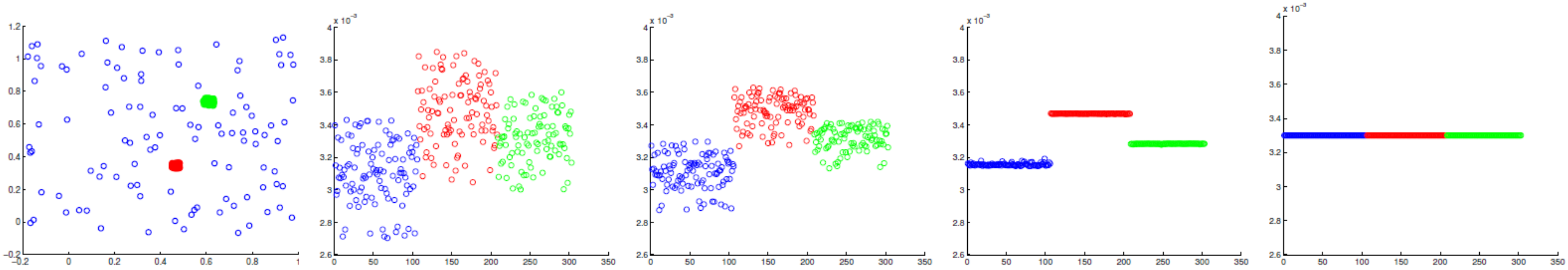




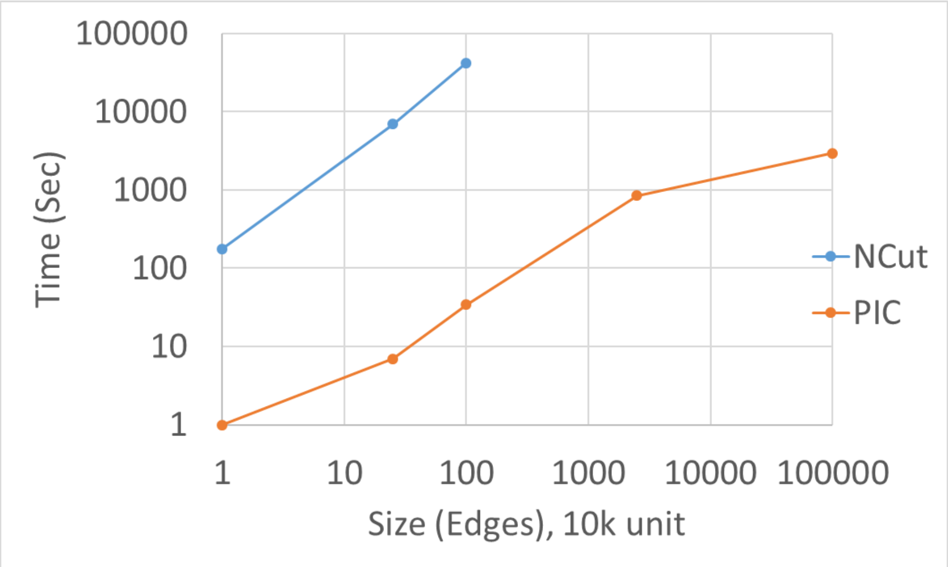
Machine Learning Components



Power-iteration-based Method



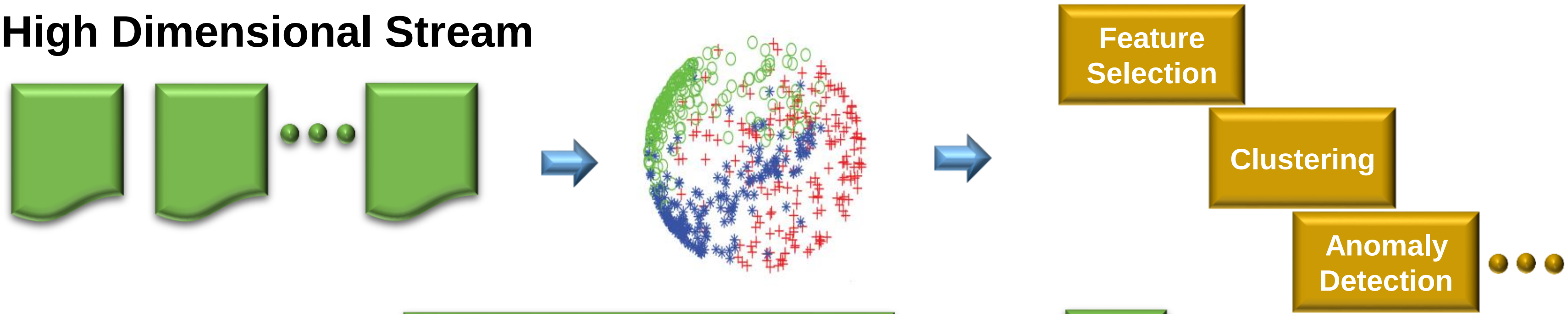
(a) 2D dataset (b) $t = 1$ (c) $t = 10$ (d) $t = 20$ (e) $t = 200$



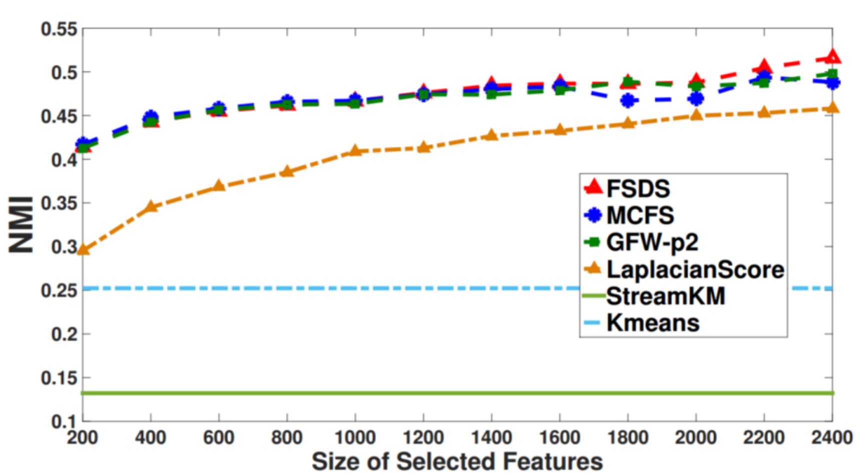
F. Lin, W. Cohen, “Power Iteration Clustering”, (ICML 2010)

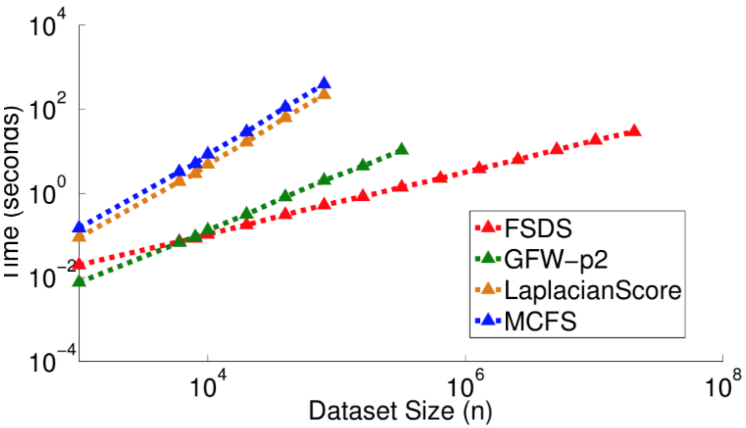
Streaming Approximations

High Dimensional Stream



Feature Selection

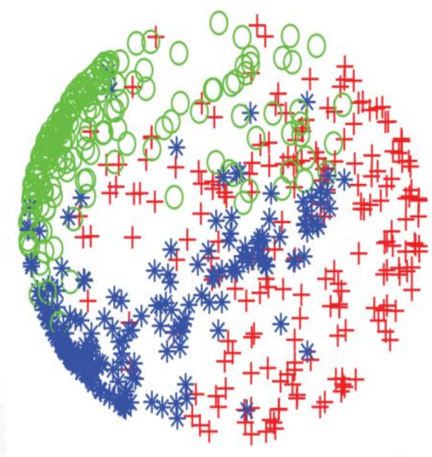


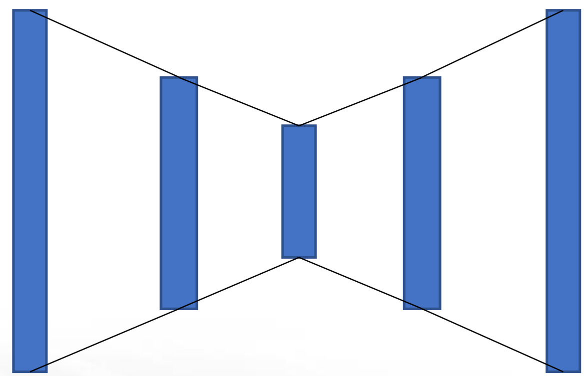


Potential Research Areas in Machine Learning

Potential Research Areas

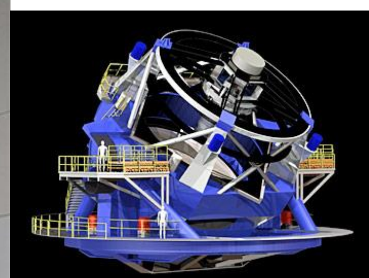
- Unsupervised / Active Learning
  - Large portion of scientific data does not have labelled data
  - “Unsupervised learning had a catalytic effect in reviving interest in deep learning, but has since been overshadowed by the successes of purely supervised learning. … we expect unsupervised learning to become far more important in the longer term.” Yann LeCun, *Nature* 2015

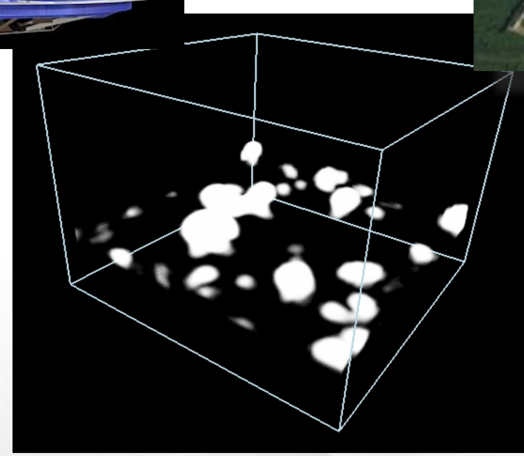




Potential Research Areas

- In-situ and streaming analysis
  - Unique much higher velocity than industry
  - Large scale simulations / cutting edge instrumentations

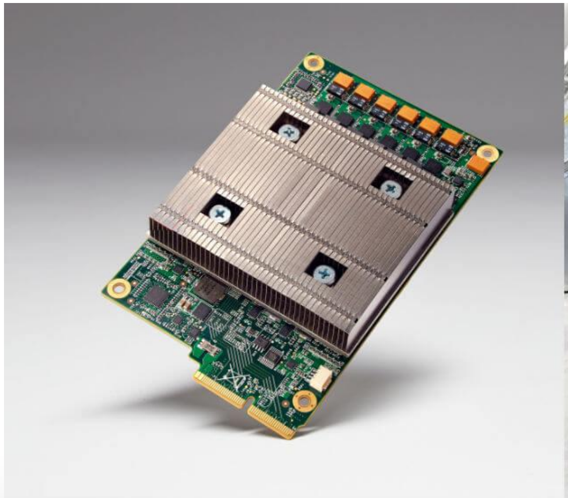




Long Island Solar Farm

Potential Research Areas

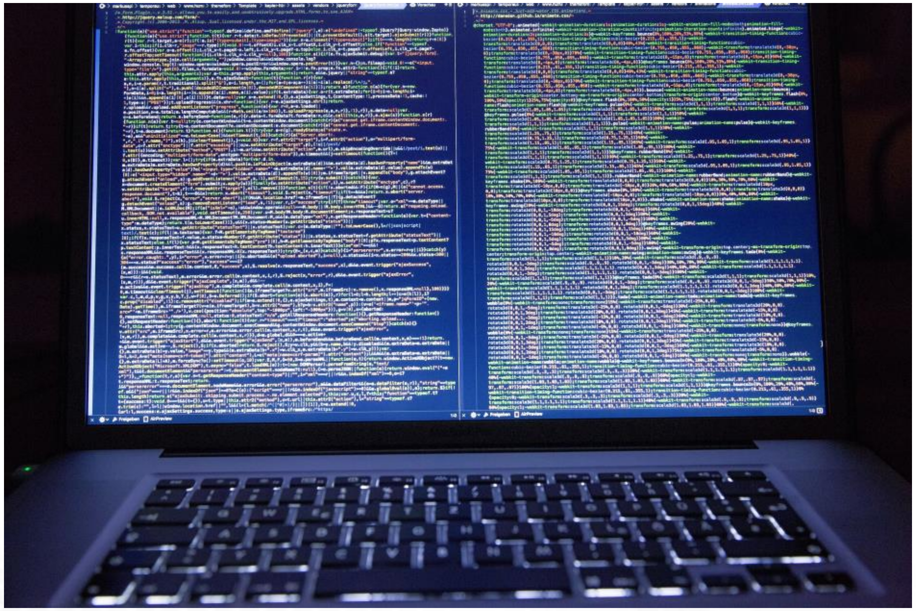
- New architectures
  - Googles' TPU (Tensor Processing Unit)
  - IBM TrueNorth (Neuromorphic Computing)



- https://futuristech.info/posts/google-claims-its-tensor-processing-unit-tpu-is-7-years-into-the-future-ahead-of-moore-s-law
- http://www.research.ibm.com/articles/brain-chip.shtml

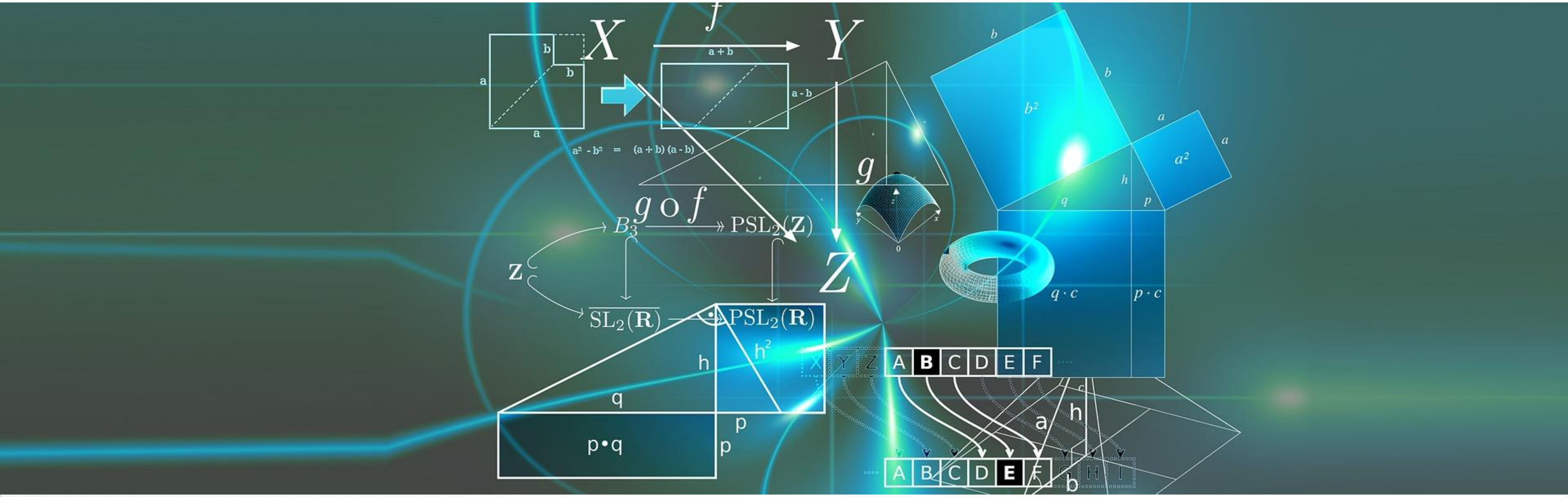
Potential Research Areas

- Programming models, compiler technologies, workflows to leverage HPC more effectively
  - Lua, Scala, Julia are popular new programming languages for machine learning



Potential Research Areas

- New mathematical solutions/solvers/libraries for HPC



Potential Research Areas

- Foundational theory for deep learning

Deep Learning without Poor Local Minima

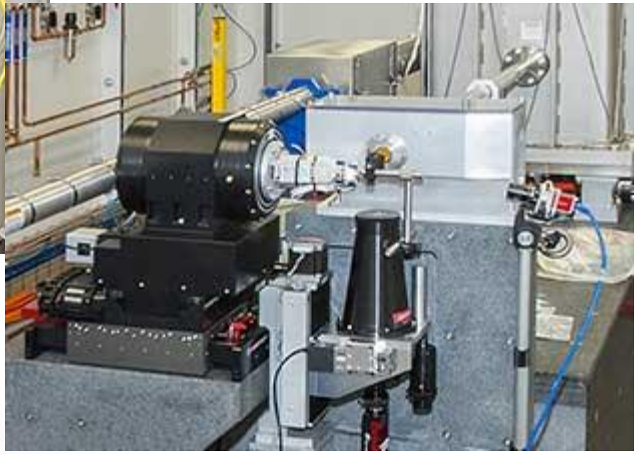
Kenji Kawaguchi
Massachusetts Institute of Technology
kawaguch@mit.edu

Abstract

In this paper, we prove a conjecture published in 1989 and also partially address an open problem announced at the Conference on Learning Theory (COLT) 2015. With no unrealistic assumption, we first prove the following statements for the squared loss function of deep linear neural networks with any depth and any widths: 1) the function is non-convex and non-concave, 2) every local minimum is a global minimum, 3) every critical point that is not a global minimum is a saddle point, and 4) there exist "bad" saddle points (where the Hessian has no negative eigenvalue) for the deeper networks (with more than three layers), whereas there is no bad saddle point for the shallow networks (with three layers). Moreover, for

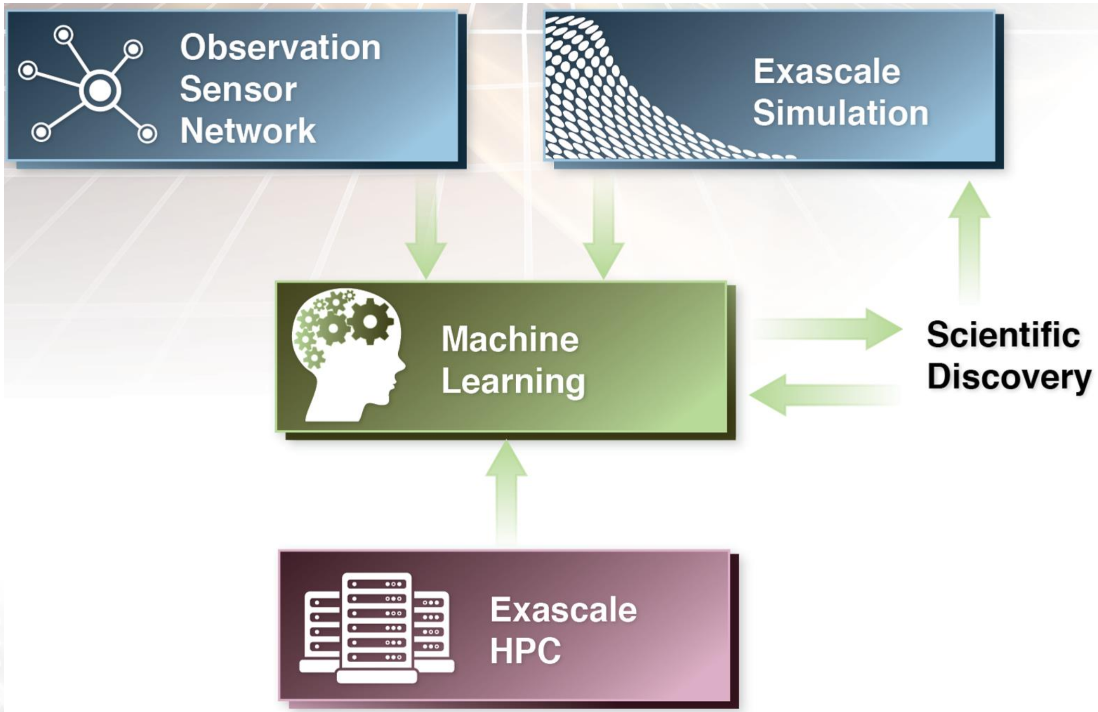
Potential Research Areas

- Automation of simulation or experiments
    - Self-driving car
    - Why not autonomous experimentation?



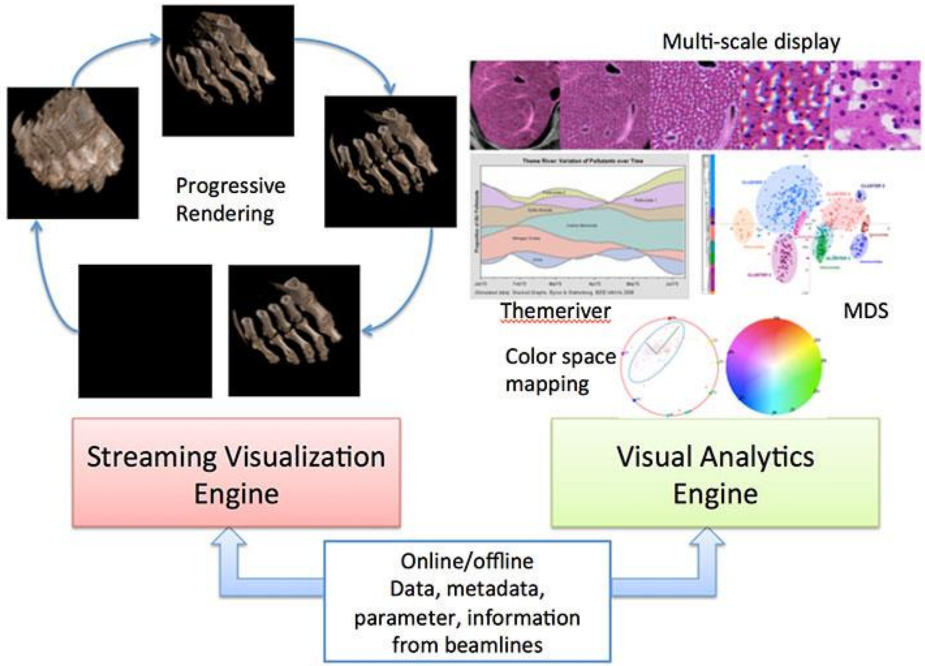
Potential Research Areas

- Fusing theory, simulation, experiments, and ML
  - Interplay of simulation, observation and ML

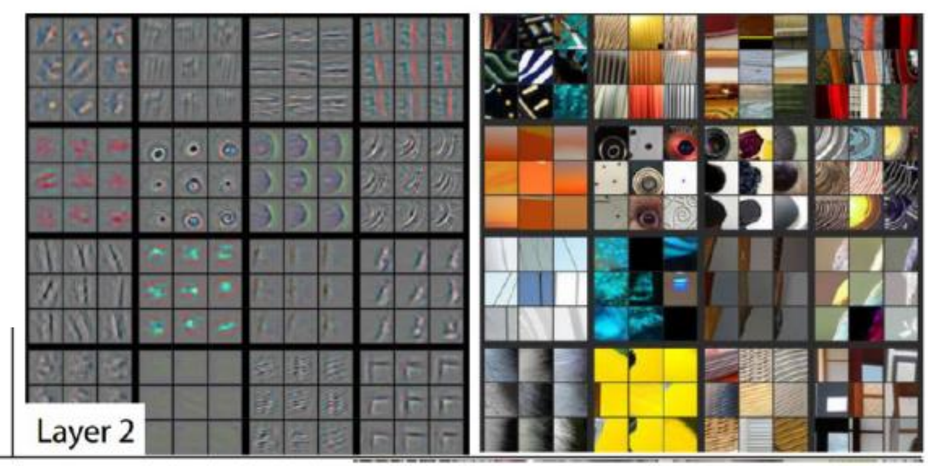


Potential Research Areas

- Interactive analysis in PB scale data
  - Enabling high dimensional feature space and high volume visualization
  - Pin-point where to pay attention
  - Good summarization and dynamic zoom-in and out
  - Help us to understand and design better machine learning algorithms

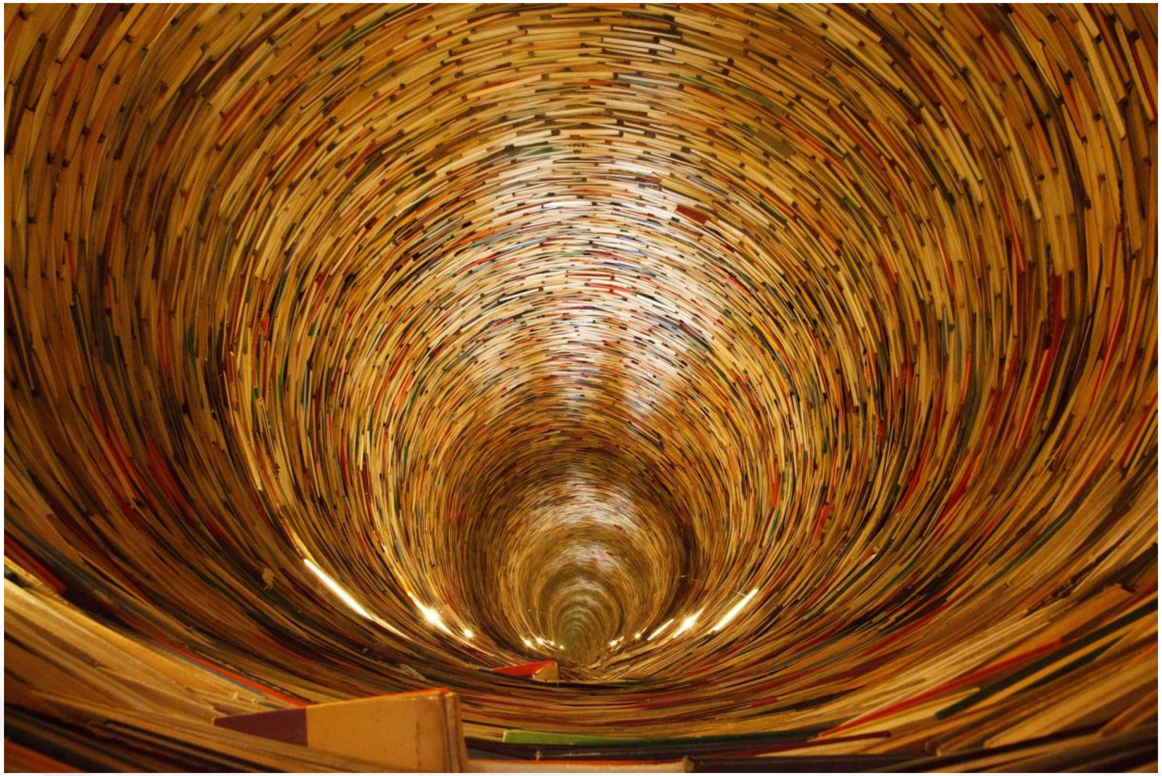


Detailed display of the individual elements in layers: here the top 9 activations of the feature maps and their corresponding image patches are plotted.

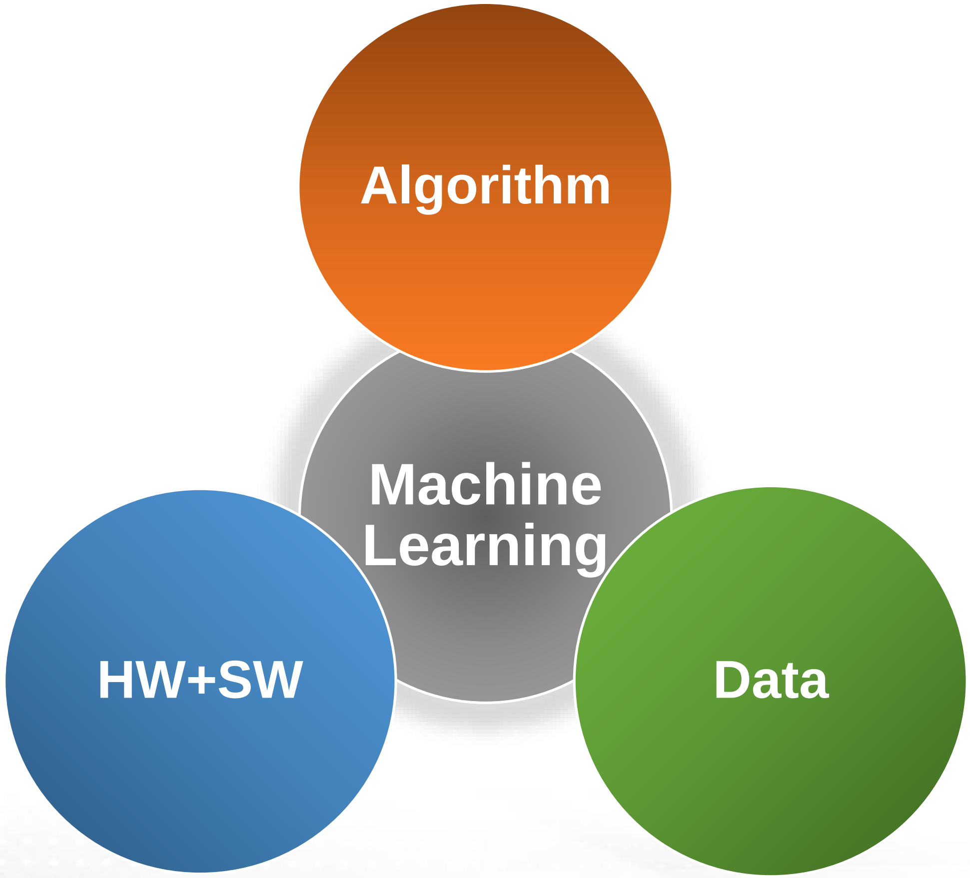


Potential Research Areas

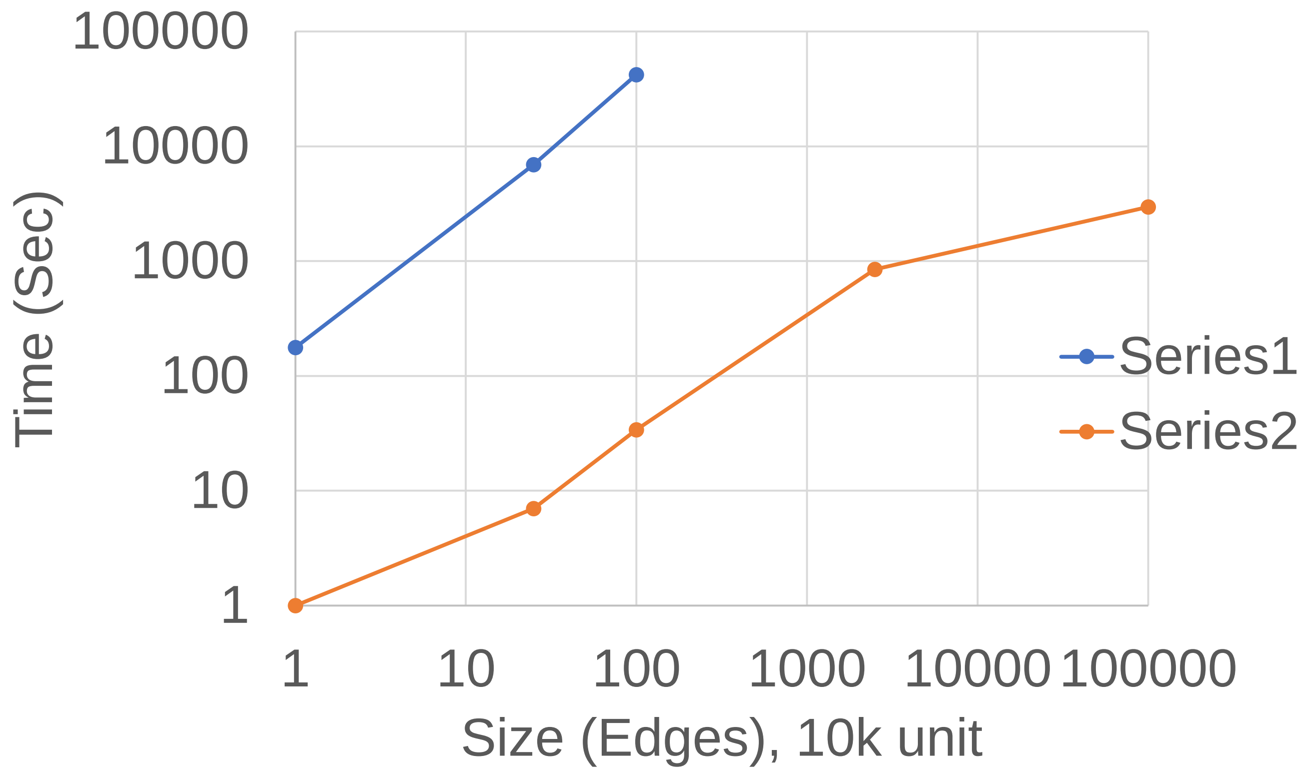
- Text Mining
  - Scientific literature was effectively utilized in various science domains



Questions?

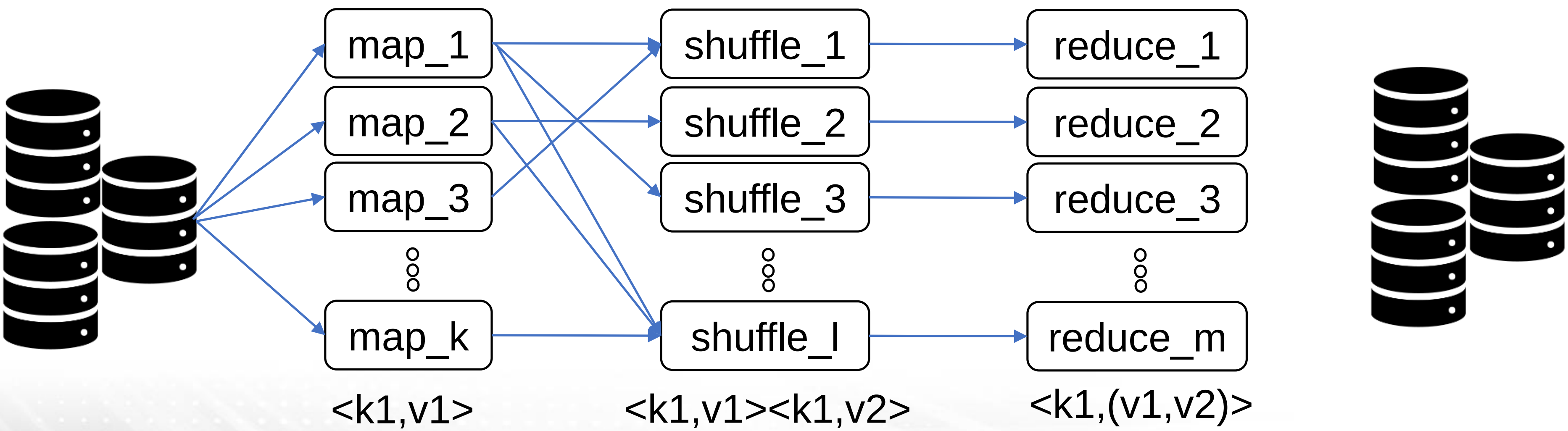


Backup


100000
10000
1000
100
10
1
Time (Sec)
1
10
100
1000
10000
100000
Size (Edges), 10k unit
Series1
Series2

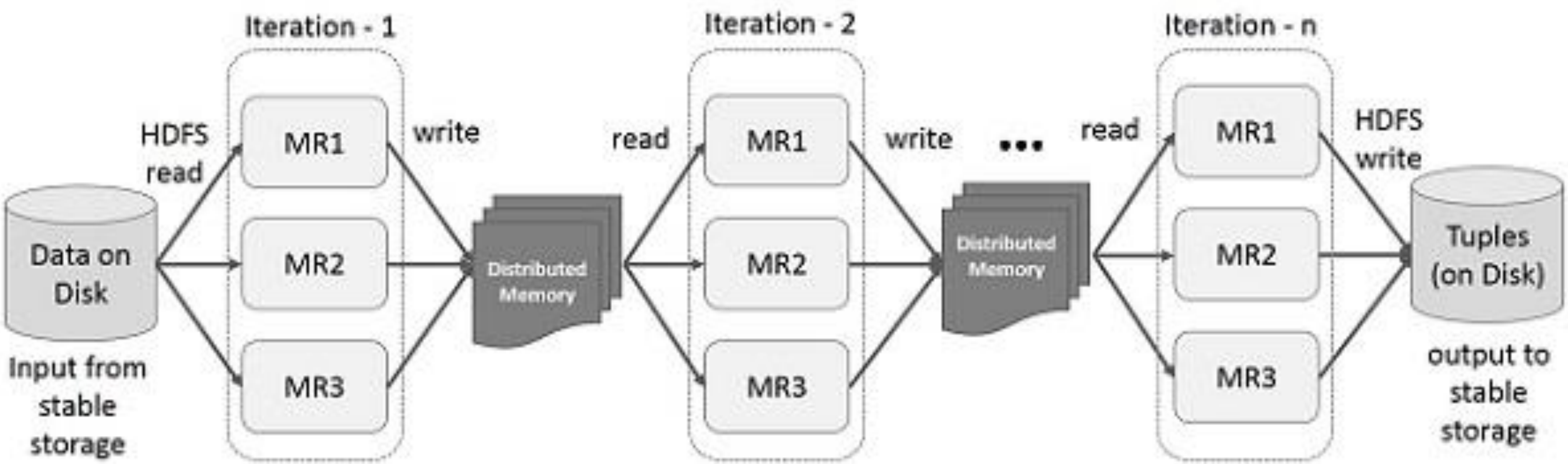
Big Data and ML

- MapReduce
  - Needed distributed processing paradigm for big volume of WWW data
  - Focused on minimizing disk IO

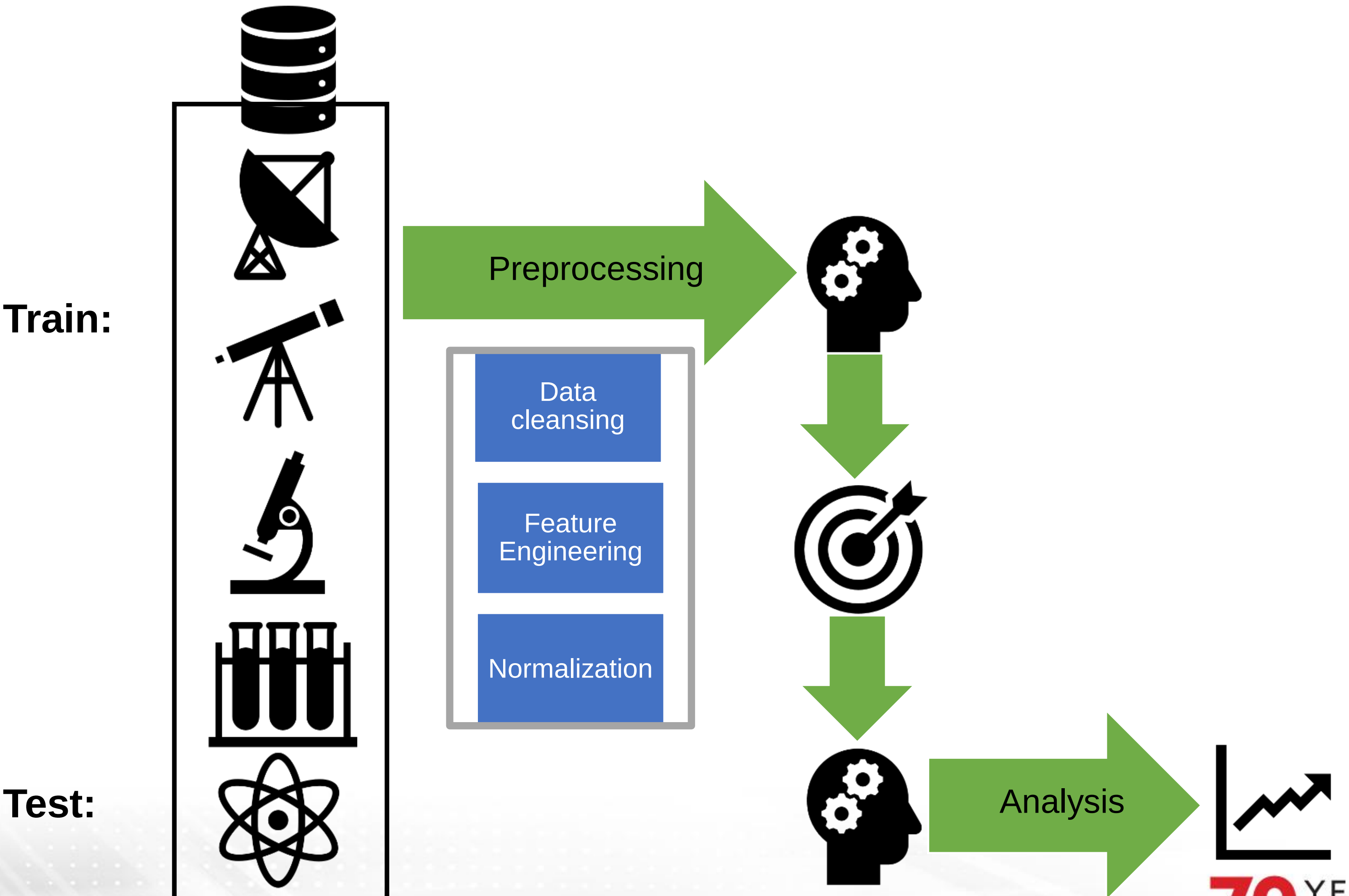


Big Data and ML

- Spark
  - Maximally utilize distributed memory (RDD)
  - Allow lazy evaluation for better optimization

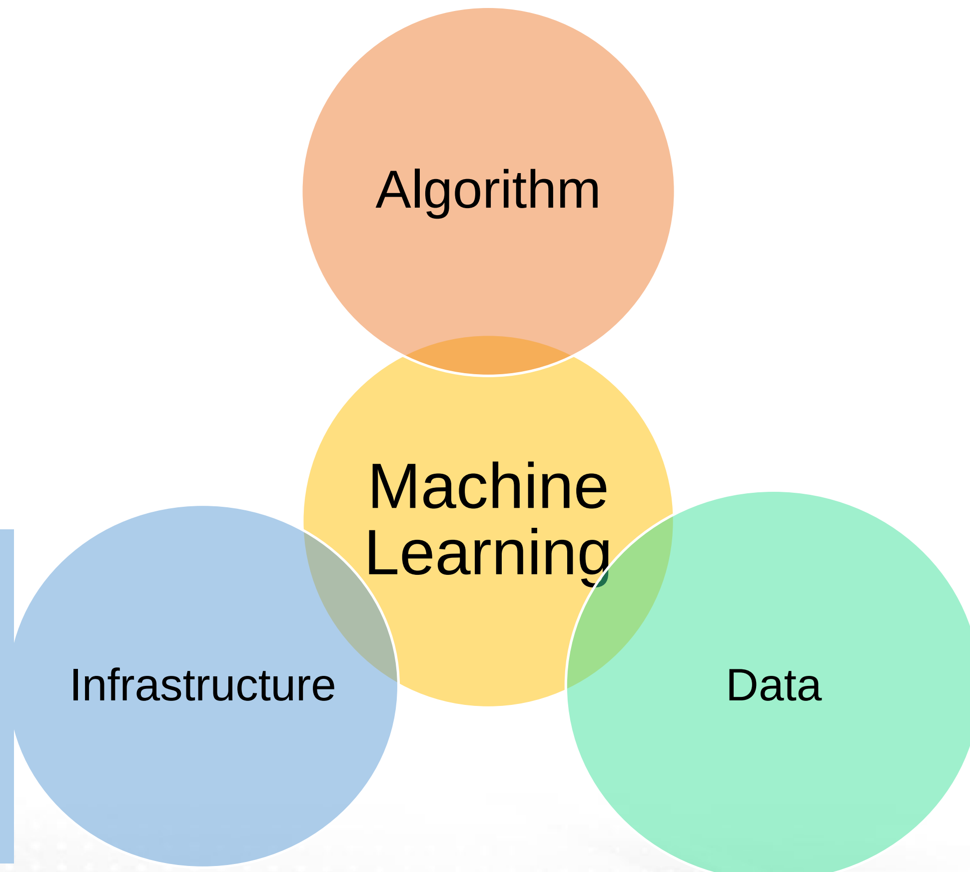


Unsupervised Learning Pipeline



Machine Learning Component

- HW: CPU, GPGPU, FPGA, …
- SW: XXX, Hadoop, spark, allreduce, …



Potential Research Areas

- Unsupervised / Active Learning
  - Large portion of scientific data does not have labelled data
  - “Unsupervised learning had a catalytic effect in reviving interest in deep learning, but has since been overshadowed by the successes of purely supervised learning. … we expect unsupervised learning to become far more important in the longer term.” Yann LeCun, *Nature* 2015
- In-situ and streaming analysis
  - Unique much higher velocity than industry
  - Large scale simulations / cutting edge instrumentations
- Programing models to leverage HPC more effectively
  - Lua, Scala, Julia are popular new programming languages for machine learning
- New architectures
  - Googles’ TPU (Tensor Processing Unit)
  - IBM TrueNorth (Neuromorphic Computing)

Potential Research Areas

- New mathematical solutions/solvers/libraries @ HPC
- Foundational theory for deep learning
- Automation of simulation or experiments
  - Self-driving car
  - Why not autonomous experimentation?
- Fusing theory, experiments, and ML
  - Interplay of simulation, observation and ML

Potential Research Areas

- Interactive analysis in PB scale data
  - Interpretable compression
  - Pin-point where to pay attention
  - Good summarization and dynamic zoom-in & out
- Text Mining
  - Scientific literature was effectively utilized in various science domain
- Error Analysis